SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0001 $01.25/0

ON THE FUNDAMENTAL THEOREM OF AVERAGING*

J. A. SANDERS†

Abstract. In this paper we present a simplified proof of the validity of first and second order averaging in the general case; the periodic case follows as a corollary.

Introduction. Although the fundamental theorem of averaging has been around for a while (Bogoliubov and Mitropolsky (1961)), and several proofs have been given besides the original (Besjes (1969) and Eckhaus (1975)), none of these proofs can be considered as very clear. Also, it has always been necessary to separate the general case from the periodic one, in order not to lose accuracy in this special, but very important, case.

In this paper we aim to give a unified approach, based mainly on Eckhaus' local averaging method, and to derive the second order approximation theory, which seems never to have been done (Van der Burgh (1974)). As a corollary we obtain an improved estimate for the first order approximation under an additional differentiability condition on the vectorfield. Since the method of proof lies between Besjes' and Eckhaus', we give rather explicit estimates in the lemmas, in order not to burden the reader with references to proofs with slightly different results and notation. Finally, in §4, we give the original proof of Bogoliubov and Mitropolsky in our notation, so the reader can easily compare the two methods.

1. On the concept of local average. In this section we shall give some definitions and lemmas found in Eckhaus (1975) with a different notation.

DEFINITION 1. Consider a function f: $\mathbb{R}\times\mathbb{R}^p\to\mathbb{R}^n$. The *local average* f_T of f is defined by

$$f_T(t,x)=\frac{1}{T}\int_0^T f(t+\tau,x)\,d\tau$$

(x is a dummy in this definition, p might be zero).

Remark. If f is periodic in t, with period T, then f_T equals the usual average f^0 where

$$f^0(x)=\frac{1}{T}\int_0^T f(t,x)\,dt.$$

DEFINITION 2. Consider the differential equation

$$\dot{x}=\varepsilon f(t,x),\qquad x\in D\subset\mathbb{R}^n.$$

Suppose f is continuous in t and x on $\mathbb{R}\times D$, and uniformly bounded (with constant M), and uniformly Lipschitz with respect to x (with constant λ, i.e., $\|f(x)-f(y)\|\le\lambda\|x-y\|$ for all $x,y\in D$). Furthermore, suppose that its *average*

$$f^0(x)=\lim_{T\to\infty}\frac{1}{T}\int_0^T f(t,x)\,dt$$

exists uniformly. Then we call f a *KBM-vectorfield.*

Remark. In the sequel we will assume that f is always a KBM-vectorfield with bound M and Lipschitz constant λ. We will not repeat these conditions in the lemmas.

*Received by the editors December 9, 1980, and in revised form June 2, 1981.
†Wiskundig Seminarium, Vrije Universiteit, Amsterdam, The Netherlands.

Remark. In the sequel we shall take T such that $\varepsilon T = o(1)$, where ε is a small parameter to be introduced. We do not exclude the case $T = O(1)$.

Outline of the proof of Theorem 1. We consider the initial value problem

$$\dot{x} = \varepsilon f(t,x), \qquad x(0) = \xi \in D \subset \mathbb{R}^n.$$

We want to compare its solution to that of the averaged system

$$\dot{z} = \varepsilon f^0(z), \qquad z(0) = \xi \in D^0 \subset D \subset \mathbb{R}^n.$$

In several earlier proofs this has been done, either directly or indirectly, by studying the locally averaged equation

$$\dot{y} = \varepsilon f_T(t,y), \qquad y(0) = \xi.$$

First we want to show that x and y are approximate. We can write x as

$$x(t) = \xi + \varepsilon \int_0^t f(\tau, x(\tau))\, d\tau$$

and y as

$$y(t) = \xi + \varepsilon \int_0^t f_T(\tau, y(\tau))\, d\tau.$$

We are therefore led to consider the function

$$\phi(t) = \int_0^t f(\tau, x(\tau))\, d\tau.$$

In Lemma 1, we show that $\phi(t) = \phi_T(t) + O(T)$, and in Lemma 2 that

$$\phi_T(t) = \int_0^t f_T(\tau, x(\tau))\, d\tau + O(T).$$

At that point, we can write x as

$$x(t) = \xi + \varepsilon \int_0^t f_T(\tau, x(\tau))\, d\tau + O(\varepsilon T).$$

This is close enough to the expression for y to estimate the difference between x and y by standard methods on a time interval of length $\frac{1}{\varepsilon}$ in Lemma 3. To show that z is an approximation to y, we first show in Lemma 4 that

$$f_T(t,x) = f^0(x) + O\left(\frac{\delta(\varepsilon)}{\varepsilon T}\right), \qquad 0 \le \varepsilon t \le L$$

where

$$\delta(\varepsilon) = \sup_{x \in D} \sup_{t \in [0, L/\varepsilon]} \varepsilon \left| \int_0^t \left[f(t,x) - f^0(x) \right] d\tau \right|.$$

One might wonder what the ε does in an expression only involving f_T and f^0, but it enters via the time scale $\frac{1}{\varepsilon}$. This means that we can write

$$y(t) = \xi + \varepsilon \int_0^t f^0(y(\tau))\, d\tau + O\left(\frac{\varepsilon \delta(\varepsilon) t}{\varepsilon T}\right),$$

and this is close enough to the formula for z to give a standard estimate in Lemma 5. The proof of the theorem then consists of applying the triangle inequality and picking the right T.

LEMMA 1. *Let ϕ be a Lipschitz-continuous map from $\mathbb{R}$ to $\mathbb{R}^n$. Then $\phi(t) = \phi_T(t) + O(T)$.*

Proof.

$$|\phi(t)-\phi_T(t)|=\left|\frac{1}{T}\int_0^T[\phi(t)-\phi(t+\tau)]\,d\tau\right|\leq\frac{1}{T}\int_0^T\lambda\tau\,d\tau=O(T).$$

where λ is the Lipschitz constant of ϕ. □

LEMMA 2. *Let x be a solution of*

$$\dot{x}=\varepsilon f(t,x),\qquad x\in D\subset\mathbb{R}^n,$$

and define

$$\phi(t)=\int_0^t f(\tau,x(\tau))\,d\tau.$$

Then, on $0\leq\varepsilon t\leq L$,

$$\phi_T(t)=\int_0^t f_T(\tau,x(\tau))\,d\tau+O(T).$$

Proof.

$$\begin{aligned}\phi_T(t)&=\frac{1}{T}\int_0^T\int_0^{t+\tau}f(\sigma,x(\sigma))\,d\sigma\,d\tau\\&=\frac{1}{T}\int_0^T\int_\tau^{t+\tau}f(\sigma,x(\sigma))\,d\sigma\,d\tau+R_1\\&=\frac{1}{T}\int_0^T\int_0^t f(\sigma+\tau,x(\sigma+\tau))\,d\sigma\,d\tau+R_1\\&=\frac{1}{T}\int_0^T\int_0^t f(\sigma+\tau,x(\sigma))\,d\tau\,d\sigma+R_1+R_2\\&=\int_0^t f_T(\tau,x(\tau))\,d\tau+R_1+R_2,\end{aligned}$$

where

$$R_1=\frac{1}{T}\int_0^T\int_0^\tau f(\sigma,x(\sigma))\,d\sigma\,d\tau=O(T).$$

Since $|f(t,x)|\leq M$ for $0\leq\varepsilon t\leq L$ and $x\in D\subset\mathbb{R}^n$,

$$R_2=\frac{1}{T}\int_0^t\int_0^T[f(\sigma+\tau,x(\sigma+\tau))-f(\sigma+\tau,x(\sigma))]\,d\tau\,d\sigma$$

and

$$\begin{aligned}|R_2|&\leq\frac{1}{T}\int_0^t\int_0^T\lambda\|x(\sigma+\tau)-x(\sigma)\|\,d\tau\,d\sigma\\&\leq\frac{\varepsilon}{T}\int_0^t\int_0^T\lambda\int_\sigma^{\sigma+\tau}|f(\sigma',x(\sigma'))|\,d\sigma'\,d\tau\,d\sigma\\&\leq\frac{\varepsilon}{T}\int_0^t\int_0^T\lambda M\tau\,d\tau\,d\sigma=O(T)\quad\text{on }0\leq\varepsilon t\leq L.\end{aligned}$$

LEMMA 3. *Let x be the solution of*

$$\dot{x}=\varepsilon f(t,x),\qquad x(0)=\xi,$$

and let y be the solution of

$$\dot{y}=\varepsilon f_T(t,y),\qquad y(0)=\xi.$$

Then $x(t)=y(t)+O(\varepsilon T)$ on $0\leq\varepsilon t\leq L$.

Proof.

$$x(t)=\xi+\varepsilon\int_0^t f(\tau,x(\tau))\,d\tau.$$

According to Lemmas 1 and 2, we know that

$$\phi(t)=\int_0^t f(\tau,x(\tau))\,d\tau=\phi_T(t)+O(T)=\int_0^t f_T(\tau,x(\tau))\,d\tau+O(T)$$

or

$$x(t)=\xi+\varepsilon\int_0^t f_T(\tau,x(\tau))\,d\tau+O(\varepsilon T).$$

Since

$$y(t)=\xi+\varepsilon\int_0^t f_T(\tau,y(\tau))\,d\tau,$$

it is readily seen, since f_T inherits the Lipschitz continuity from f, that

$$x(t)=y(t)+O(\varepsilon T).$$ □

Remark. In case f is periodic, this proves the fundamental theorem of averaging, since $f_T=f^0$ then.

LEMMA 4. *The local average and the average of f are related by the estimate*

$$f_T(t,x)=f^0(x)+O\left(\frac{\delta(\varepsilon)}{\varepsilon T}\right),\qquad 0\le\varepsilon t\le L,$$

where

$$\delta(\varepsilon)=\sup_{x\in D}\sup_{t\in[0,L/\varepsilon)}\varepsilon\left|\int_0^t\left[f(\tau,x)-f^0(x)\right]d\tau\right|.$$

Proof.

$$\begin{aligned}f_T(t,x)-f^0(x)&=\frac{1}{T}\int_0^T\left[f(\tau+t,x)-f^0(x)\right]d\tau\\&=\frac{1}{T}\int_0^{T+\tau}\left[f(\tau,x)-f^0(x)\right]d\tau+\frac{1}{T}\int_0^t\left[f(\tau,x)-f^0(x)\right]d\tau.\end{aligned}$$

Since for $\alpha=o(t)$,

$$\left|\frac{1}{T}\int_0^{t+\alpha}\left[f(\tau,x)-f^0(x)\right]d\tau\right|\le\frac{\delta(\varepsilon)}{\varepsilon T};$$

this gives the desired result. □

LEMMA 5. *Let y be the solution of*

$$\dot y=\varepsilon f_T(t,y),\qquad y(0)=\xi,$$

and let z be the solution of

$$\dot z=\varepsilon f^0(z),\qquad z(0)=\xi.$$

Then

$$y(t)=z(t)+O\left(\frac{\delta(\varepsilon)}{\varepsilon T}\right)\quad \text{on } 0\le\varepsilon t\le L.$$

Proof. This follows from Lemma 4 and Gronwall's inequality. □

2. First order averaging. We are now able to formulate and "prove" (there is not much work left) the classical averaging theorem.

THEOREM 1 (fundamental theorem of averaging). *Consider the solution x of*

$$\dot{x}=\varepsilon f(t,x), \quad x(0)=\xi, \quad x\in D\subset\mathbb{R}^n.$$

and z, which is a solution of

$$\dot{z}=\varepsilon f^0(z), \quad z(0)=\xi, \quad z\in D^0\subset D\subset\mathbb{R}^n.$$

(We assume f to be a KBM-vectorfield, as defined in §1.) Suppose f is Lipschitz continuous on D and let L be such that $z(t)\in D^0$ for all t such that $0\leq\varepsilon t\leq L$, where L is independent of ε. Suppose furthermore that the boundaries of D and D^0 have $O_S(1)$-distance (i.e., $O(1)$, but not $o(1)$). Then

$$x(t)=z(t)+O\left(\sqrt{\delta_1(\varepsilon)}\right).$$

where

$$\delta_1(\varepsilon)=\sup_{x\in D}\sup_{t\in[0,L/\varepsilon)}\varepsilon\left|\int_0^t\left[f(\tau,x)-f^0(x)\right]d\tau\right|,$$

$$f^0(x)=\lim_{T\to\infty}\frac{1}{T}\int_0^T f(t,x)\,dt.$$

If f is periodic in t, then

$$x(t)=z(t)+O(\varepsilon).$$

Proof. The periodic case follows from Lemma 3.

In the general case we combine Lemma 3 and Lemma 5 to obtain

$$x(t)=z(t)+O(\varepsilon T)+O\left(\frac{\delta_1(\varepsilon)}{\varepsilon T}\right).$$

If we let

$$\varepsilon^2T^2=\delta_1(\varepsilon),$$

we get

$$x(t)=z(t)+O\left(\sqrt{\delta_1(\varepsilon)}\right).$$

This choice of T is in accordance with the requirement that εT be $o(1)$ (unless, of course, $\delta_1(\varepsilon)$ is not $o(1)$, in which case the result is worthless anyway).

The reader should check that the requirement $z(t)\in D^0$ has been used implicitly in the proof of the lemmas. To do everything right, one could use continuous induction on t. Since the distance of the boundaries is $O_S(1)$ and the approximation is $o(1)$, this should give no difficulties. □

Remark. If f is a finite sum of periodic functions with different periods, one can still obtain $O(\varepsilon)$-accuracy, due to the linearity of the argument leading to the estimates. In Eckhaus' proof (1975) this argument may not work, since use is made of x_T, and it is not clear how to generalize this to more periods. As we showed, the use of x_T is not necessary.

3. Second order averaging. We shall now turn to higher order approximations. In the periodic case, this is a well-established theory with many applications, but we do not know of any results in this direction in the general case. This might be due to a lack of practical importance, but, on the other hand, that argument never stopped a

mathematician. From the higher order argument it does follow, however, that the first order approximation is better than we expected. We shall prove it to have $O(\delta_1(\varepsilon))$-accuracy, under an additional differentiability hypothesis. We shall follow closely the structure of the periodic theory, so all our results are the best possible in that special case. In this section we will make the following assumption: f is a KBM-vectorfield with a uniformly Lipschitz continuous first derivative (in x).

LEMMA 6. *Let x be the solution of*

$$\dot{x}=\varepsilon f(t,x), \qquad x(0)=\xi.$$

Let w be defined by

$$x(t)=w(t)+\delta_1(\varepsilon)u^1(t,w(t)),$$

where

$$\delta_1(\varepsilon)u^1(t,w)=\varepsilon\int_0^t\left[f(\tau,w)-f^0(w)\right]d\tau.$$

(*Clearly, u^1 is uniformly bounded on $0\leq\varepsilon t\leq L$.*)

Then

$$\begin{aligned}w(t)=\xi&+\varepsilon\int_0^t f^0(w(\tau))\,d\tau\\&+\varepsilon\delta_1(\varepsilon)\int_0^t\left[\nabla f(\tau,w(\tau))u^1(\tau,w(\tau))-\nabla u^1(\tau,w(\tau))f^0(w(\tau))\right]d\tau\\&+O\left(\delta_1^2\right)\quad \text{on } 0\leq\varepsilon t\leq L.\end{aligned}$$

Proof. This is a standard computation. We use the fact that

$$\frac{du^1}{dt}(t,w(t))=\frac{\partial u^1}{\partial t}+\nabla u^1\cdot\frac{dw}{dt}.$$

$$\begin{aligned}w(t)&=x(t)-\delta_1(\varepsilon)u^1(t,w(t))\\&=\xi+\varepsilon\int_0^t f(\tau,x(\tau))\,d\tau-\varepsilon\int_0^t\left[f(\tau,w(\tau))-f^0(w(\tau))\right]d\tau\\&\quad-\delta_1(\varepsilon)\int_0^t\nabla u^1(\tau,w(\tau))\,\frac{dw}{dt}\,d\tau\\&=\xi+\varepsilon\int_0^t f^0(w(\tau))\,d\tau\\&\quad+\varepsilon\delta_1(\varepsilon)\int_0^t\left[\nabla f(\tau,w(\tau))u^1(\tau,w(\tau))-\nabla u^1(\tau,w(\tau))f^0(w(\tau))\right]d\tau+O\left(\delta_1^2\right).\end{aligned}$$

□

LEMMA 7. *Let w be as in Lemma 6 and let v be the solution of*

$$\dot{v}=\varepsilon f^0(v)+\varepsilon\delta_1(\varepsilon)f_T^1(t,v)$$

where

$$f^1(t,v)=\nabla f(t,v)u^1(t,v)-\nabla u^1(t,v)f^0(v).$$

Then

$$w(t)=v(t)+O\left(\delta_1(\varepsilon)\left(\varepsilon T+\delta_1(\varepsilon)\right)\right).$$

Proof. See Lemma 3. □

LEMMA 8. *Suppose f and f^1 (as defined in Lemma 7) are KBM-vectorfields. Let u be the solution of*

$$\dot{u}=\varepsilon f^0(u)+\varepsilon\delta_1(\varepsilon)f^{10}(u), \qquad u(0)=\xi$$

(where f^{10} is the average of f^1). Then if

$$\delta_2(\varepsilon)=\sup_{x\in D}\sup_{t\in[0,L/\varepsilon)}\varepsilon\left|\int_0^t f^1(\tau,x)-f^{10}(x)\,d\tau\right|,$$

we have

$$v(t)=u(t)+O\left(\frac{\delta_1(\varepsilon)\delta_2(\varepsilon)}{\varepsilon T}\right).$$

Proof. See Lemma 5. □

THEOREM 2 (second order approximation). *Under the hypotheses of Theorem 1, together with those of Lemma 6, we have*

$$x(t)=u(t)+\delta_1(\varepsilon)u^1(t,u(t))=O\left(\delta_1(\varepsilon)\left(\sqrt{\delta_2}+\delta_1\right)\right)$$

where u is the solution of

$$\dot{u}=\varepsilon f^0(u)+\varepsilon\delta_1(\varepsilon)f^{10}(u), \qquad u(0)=\xi,$$

and $f^{10}(u)=\lim_{T\to\infty}\frac{1}{T}\int_0^T[\nabla f(t,u)u^1(t,u)-\nabla u^1(t,u)f^0(u)]\,dt$.

Proof. See Theorem 1. □

THEOREM 3 (improved first order). *Under the hypotheses of Theorem 2, we find*

$$x(t)=z(t)+O(\delta_1(\varepsilon))$$

where z is the solution of

$$\dot{z}=\varepsilon f^0(z), \qquad z(0)=\xi.$$

Proof. Evidently,

$$u(t)=z(t)+O(\delta_1) \quad \text{on } 0\le\varepsilon t\le L.$$

Since

$$x(t)=u(t)+O(\delta_1(\varepsilon)) \quad \text{on } 0\le\varepsilon t\le L,$$

we have

$$x(t)=z(t)+O(\delta_1),$$

the desired estimate. □

4. Alternative estimate for the fundamental theorem. We have by now seen two different approaches to first order averaging: direct estimation of the differential equations (§2) and using a transformation (§3), the latter method giving better estimates but requiring differentiability of the vectorfield. In this section, we shall translate the original version of the proof of Bojoliubov and Mitropolsky into our notation. The original proof is more concerned with continuity than with asympotic estimates, but the translation is straightforward, well known and not published. The only change made here is that we allow for a bounded domain; this introduces minor technical difficulties.

The idea of the proof is rather simple: if the transformation function u^1 is not differentiable (because the vectorfield is not), we approximate it, using convolution, by a differentiable function u^1_μ. The inherent difficulty is that the gradient of u^1_μ might be

rather large, in terms of the order of approximation. All this results in a new proof of Theorem 1, except that the accuracy in the periodic case is only $O(\sqrt{\varepsilon})$, instead of $O(\varepsilon)$.

DEFINITION 3. Let $D^0 \subset D$ be such that $\operatorname{dist}(\partial D, \partial D^0) > \mu > 0$. Let $\psi : D \to \mathbb{R}$ be a continuous function. Then we define $\psi_\mu : D^0 \to \mathbb{R}$ as

$$\psi_\mu(x) = \int_D \Phi^\mu(x-y)\psi(y)\,dy,$$

where

$$\Phi^\mu(x) = \begin{cases} A_\mu\left(1 - \dfrac{\|x\|^2}{\mu^2}\right), & \|x\| \le \mu, \\ 0, & \|x\| > \mu \end{cases}$$

with A_μ such that $\int \Phi^\mu(x)\,dx = 1$.

LEMMA 9.

$$\int \nabla \Phi^\mu(x)\,dx = O\left(\frac{1}{\mu}\right).$$

Proof. Straightforward computation. □

LEMMA 10. *If $\psi : D \to \mathbb{R}$ is uniformly bounded and Lipschitz continuous, i.e., $\|\psi(x) - \psi(y)\| \le \lambda \|x-y\|$ for all $x, y \in D$, then*

$$\psi(x) = \psi_\mu(x) + O(\mu)$$

uniformly on D^0.

Proof. Let $x \in D^0$. Then

$$\begin{aligned} |\psi(x) - \psi_\mu(x)| &= \left|\psi(x) - \int \Phi^\mu(x-y)\psi(y)\,dy\right| \\ &= \left|\psi(x)\int \Phi^\mu(x-y)\,dy - \int \Phi^\mu(x-y)\psi(y)\,dy\right| \\ &= \left|\int \Phi^\mu(x-y)(\psi(x) - \psi(y))\,dy\right| \\ &\le \int |\Phi^\mu(x-y)|\lambda|\,|x-y|\,|\,dy \le \lambda\mu. \end{aligned}$$

□

LEMMA 11. *Suppose f is Lipschitz continuous with respect to x. Let x be the solution of*

$$\dot{x} = \varepsilon f(t,x), \qquad x(0) = \xi.$$

Let w be defined by

$$x(t) = w(t) + \delta_1(\varepsilon)u^1(t, w(t))$$

where

$$\delta_1(\varepsilon)u^1(t,w) = \varepsilon \int_0^t [f(\tau, w) - f(w)]\,d\tau.$$

Then

$$w(t)=\xi+\varepsilon\int_0^t f^0(w(\tau))\,d\tau+O\left(\sqrt{\delta_1(\varepsilon)}\right).$$

Proof.

$$\begin{aligned}
w(t)&=x(t)-\delta_1(\varepsilon)u^1(t,w(t))\\
&=\xi+\varepsilon\int_0^t f(\tau,x(\tau))\,d\tau-\delta_1(\varepsilon)u_\mu^1(t,w(t))+O(\delta_1\mu)\\
&=\xi+\varepsilon\int_0^t f(\tau,x(\tau))\,d\tau-\delta_1(\varepsilon)\int_0^t\frac{d}{d\tau}u_\mu^1(\tau,w(\tau))\,d\tau+O(\delta_1\mu)\\
&=\xi+\varepsilon\int_0^t f(\tau,x(\tau))\,d\tau-\varepsilon\int_0^t\left[f_\mu(\tau,w(\tau))-f_\mu^0(w(\tau))\right]d\tau\\
&\quad-\delta_1(\varepsilon)\int_0^t\nabla u_\mu^1\cdot\frac{dw}{d\tau}(\tau)\,d\tau+O(\delta_1\mu)\\
&=\xi+\varepsilon\int_0^t f^0(w(\tau))\,d\tau+\varepsilon\int_0^t\left[f_\mu^0(w(\tau))-f^0(w(\tau))\right]d\tau\\
&\quad+\varepsilon\int_0^t\left[f(\tau,x(\tau))-f_\mu(\tau,x(\tau))\right]d\tau\\
&\quad+\delta_1(\varepsilon)\int_0^t\left[\varepsilon\nabla f_\mu\cdot u_\mu^1-\nabla u_\mu^1\frac{dw}{d\tau}\right]d\tau+O(\delta_1\mu)\\
&=\xi+\varepsilon\int_0^t f^0(w(\tau))\,d\tau+O(\mu)+O\left(\frac{\delta_1}{\mu}\right).
\end{aligned}$$

So we let $\mu=\sqrt{\delta_1(\varepsilon)}$ and we obtain the desired estimate. □

Using Lemma 11 and Gronwall's lemma, we obtain a second proof of Theorem 1, except that the estimate is not very sharp in the periodic case.

Remark. If one looks over the proof of Lemma 11, one gets the distinct feeling that there must be sharper estimates. One could for instance apply the averaging theory to the function $g_\mu(t,x)=f(t,x)-f_\mu(t,x)$. The difficulty here is that, while g_μ is itself $O(\mu)$, its Lipschitz constant can only be shown to be $O(1)$, and it is also not clear that the δ-function belonging to g_μ is smaller than δ_1, even if this might seem intuitively acceptable. One of the problems here is that $\dot{x}=O(\varepsilon)$ and not $O(\varepsilon\mu)$.

5. Concluding remarks. We do not present here a general theory of higher order approximations; suffice it to say that the proof technique presented here can be easily used, once the complications of the formal analysis are understood.[1] If f^0 happens to be zero, then one can easily extend the result on the time scale $0\le\varepsilon\delta_1(\varepsilon)t\le L$, dividing of course the accuracy by δ_1. If f^0 has an attracting stationary point, solutions starting well inside the attraction domain can be approximated on $[0,\infty)$ without loss of accuracy. If f^0 has a hyperbolic point, then solutions on the stable manifold of the corresponding orbit solution of f can be also approximated by solutions on the stable manifold of f^0 with validity on $[0,\infty)$.

---

[1] To avoid notational difficulties, a multi-index notation for the small parameters should be used to analyze the formal problem.

REFERENCES

J. G. BESJES, (1969), *On the asymptotic method for non-linear differential equations*, J. Mécanique, 8, pp. 357–372.

N. N. BOGOLIUBOV AND Y. A. MITROPOLSKY, (1961) *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon & Breach, New York.

W. ECKHAUS, (1975), *New approach to the asymptotic theory of nonlinear oscillations and wave-propagation*, J. Math. Anal. Appl., 49, pp. 575–611.

A. H. P. VAN DER BURGH, (1974), *Studies in the asymptotic theory of nonlinear resonance*, Thesis, Technische Hochschule, Delft.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0002 $01.25/0

ANALYSIS OF BOUNDARY VALUE PROBLEMS ON INFINITE INTERVALS*

PETER A. MARKOWICH†

Abstract. In this paper boundary value problems of ordinary differential equations on infinite intervals are analysed. There is a theory for problems of this kind which requires the fundamental matrix of the system of differential equations to have certain decay properties near infinity. The aim of this paper is to establish a theory which holds under weaker assumptions. The analysis for linear problems is done by determining the fundamental matrix of the system of differential equations asymptotically. For inhomogeneous problems a suitable particular solution having a "nice" asymptotic behaviour is chosen and so global existence and uniqueness theorems are established in the linear case. The asymptotic behaviour of this solution follows immediately. Nonlinear problems are treated by using perturbation techniques meaning linearization near infinity and by using the methods for the linear case. Moreover, some problems from fluid dynamics and thermodynamics are dealt with and they illustrate the power of the asymptotic methods used.

Key words. nonlinear boundary value problems, singular points, boundedness, asymptotic properties, asymptotic expansions

1. Introduction. This paper is concerned with the analysis of boundary value problems on infinite intervals posed as follows:

$$y' = t^{\alpha} f(t,y), \qquad 1 \le t < \infty, \quad \alpha \in N_0, \tag{1.1}$$

$$y \in C([1,\infty]) : \Leftrightarrow y \in C([1,\infty]) \text{ and } \lim_{t\to\infty} y(t) = y(\infty) \text{ is finite}, \tag{1.2}$$

$$b(y(1), y(\infty)) = 0, \tag{1.3}$$

where y is an n-vector, f and b are nonlinear mappings.

If f does not depend on t explicitly then $\alpha \in R$ such that $\alpha > -1$ can be assumed instead of $\alpha \in N_0$.

Equation (1.1) has a singularity of the second kind at $t = \infty$ of rank $\alpha + 1$, since we assume that f is continuous in $(\infty, y(\infty))$. The goal is to establish existence and uniqueness theorems for very general f and b, to gain information on the behaviour of y for large t and—in the linear inhomogeneous case— to investigate the dependence of y on the boundary data and the inhomogeneity.

Problems of this kind frequently occur in fluid dynamics when similarity solutions of the stationary Navier–Stokes equations for certain flow-constellations are sought (see for example McLeod (1969), Markowich (1982a), Lentini and Keller (1980b), Cohen, Fokas and Lagerström (1978)).

For application in other areas of physics see Lentini (1978).

Much analytical work has been done on singular boundary value problems of the second kind. De Hoog and Weiss (1980a,b) investigated the case where $(\partial f/\partial y)(\infty, y(\infty))$ has no eigenvalue on the imaginary axis by linearizing f around $y(\infty)$ and evaluating at $t = \infty$, so getting the constant coefficient problem

$$z' = t^{\alpha} f_y(\infty, y(\infty)) z, \tag{1.4}$$

* Received by the editors November 14, 1980, and in revised form November 30, 1981. This work was sponsored 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-7927062.

† Department of Mathematics, University of Texas, Austin, Texas 78712.

and by employing perturbation techniques which are based on estimates derived for a certain particular solution of linear inhomogeneous problems. They established uniqueness and existence theorems. Lentini and Keller (1980a) extended this approach, neglecting the assumption on the eigenvalues, but they required that the projection of $f_y(t,y(\infty))$ onto the direct sum of invariant subspaces of $f_y(\infty,y(\infty))$ which correspond to an imaginary eigenvalue converge, at least as $t^{-(\alpha+1)r-\varepsilon}$, where r is the largest dimension of these subspaces and $\varepsilon>0$. It turns out that this assumption is crucial and the perturbation approach breaks down if $f_y(\infty,y(\infty))$ has eigenvalues with real part zero and if the convergence requirement is neglected. However, many physical problems do not fulfill this covergence requirement in the presence of imaginary eigenvalues (see for example Cohen, Fokas and Lagerström (1978) and Schlichting (1951)) and therefore a more general approach is necessary.

In this paper asymptotic series are used in order to determine asymptotically fundamental solution matrices of linear systems of the form:

$$y'=t^{\alpha}A(t)y, \qquad t\geq\delta. \tag{1.5}$$

The basic assumption is that A is analytic in $[\delta,\infty]$ for some $\delta\geq 1$, so that

$$A(t)=\sum_{i=0}^{\infty}A_i t^{-i} \quad \text{where } A_i=\frac{1}{i!}\lim_{x\to 0^+}\frac{d^i}{dx^i}A\left(\frac{1}{x}\right). \tag{1.6}$$

Then a fundamental matrix $\Phi(t)$ of (1.5) can be calculated as an asymptotic (formal) log-exponential series from the coefficients A_i by a recursive algorithm (see Coddington and Levinson (1955) and Wasow (1965)).

Assumption (1.6) can be weakened so that only a finite but large enough number of these derivatives exist.

Equation (1.1) is treated by linearization around $y(\infty)$ obtaining the variable coefficient problem

$$z'=t^{\alpha}f_y(t,y(\infty))z \tag{1.7}$$

and again by employing perturbation techniques.

The advantage of the formal-series approach is twofold. Firstly, no restrictions (except (1.6)) have to be made on the convergence behaviour of $f_y(t,y(\infty))$; secondly, the asymptotic behaviour of the (basic) solutions is obtained directly. The asymptotic behaviour is crucial for the determination of appropriate numerical procedures for Problems (1.1), (1.2), (1.3) (See Lentini and Keller (1980a), Markowich (1980), (1982b) and de Hoog and Weiss (1980b).)

Recently Jepson (1981) and Markowich (1982b) used the asymptotic expansions of the fundamental matrix $\phi(t)$ in order to set up finite (asymptotic) boundary conditions for the numerical solution of (1.1), (1.2), (1.3).

This paper is organized as follows. In §2 some remarks are made on linear inhomogeneous constant coefficient problems (see Lentini and Keller (1980a)), in §3 we admit variable coefficient matrices and in §4 we get to nonlinear problems of the form (1.1), (1.2), (1.3). Section 5 is concerned with physical problems which illustrate the power of the used asymptotic methods.

2. Linear problems with constant coefficients. We consider problems of the form

$$y'-t^{\alpha}Ay=t^{\alpha}f(t), \qquad 1\leq t<\infty, \quad \alpha\in R, \quad \alpha>-1, \tag{2.1}$$

$$y\in C([1,\infty]), \tag{2.2}$$

$$B_1 y(1)+B_\infty y(\infty)=\hat{\beta}, \tag{2.3}$$

where the $n\times n$ matrix $A\neq 0$.

First we transform A to its Jordan canonical form J

$$A=EJE^{-1} \tag{2.4}$$

and substitute

$$u=E^{-1}y, \tag{2.4a}$$

so we get the new problem

$$u'-t^\alpha Ju=t^\alpha E^{-1}f(t), \tag{2.5}$$

$$u\in C([1,\infty]), \tag{2.6}$$

$$B_1 Eu(1)+B_\infty Eu(\infty)=\hat{\beta}. \tag{2.7}$$

Without loss of generality we can assume that J has the block diagonal form

$$J=\operatorname{diag}(J^+,J^0,J^-), \tag{2.8}$$

where the real parts of the eigenvalues of J^+ are positive, the real parts of the eigenvalues of J^0 are equal to zero and the real parts of the eigenvalues of J^- are negative. This structure can always be obtained by reordering the columns of E. Let the dimensions of these three matrices be r_+, r_0 and r_- respectively.

The diagonal projections D_+, D_0, D_- are obtained by taking the main-diagonal of J and by replacing every eigenvalue with positive, zero or negative real part, respectively, by 1 and all others by zero so that

$$I=D_+ +D_0+D_- \tag{2.9}$$

holds.

Furthermore let $\bar{D}_0$ be the projection onto the direct sum of eigenspaces of J associated with zero eigenvalues, which is obtained by replacing by zero every (diagonal) element of D_0 which is not associated with the first column of a Jordan block of J belonging to a zero eigenvalue.

Let the number of nonzero columns of $\bar{D}_0$ which equals the geometric multiplicity of the eigenvalue zero be $\bar{r}_0$. The general solution of the homogeneous problem (2.5), (2.6) is

$$u_h(t)=\phi(t)(\bar{D}_0+D_-)\xi=\exp\left(J\frac{t^{\alpha+1}}{\alpha+1}\right)(\bar{D}_0+D_-)\xi, \qquad \xi\in C^n. \tag{2.10}$$

In order to solve the inhomogeneous problem (2.5) we look for a particular solution $u_p\in C([1,\infty])$.

De Hoog and Weiss (1980a,b) and Lentini and Keller (1980a) suggested the following choice:

$$\begin{aligned} u_p(t)=(Hf)(t)=\phi(t)\int_\infty^t D_+\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds \\ +\phi(t)\int_\infty^t D_0\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds \\ +\phi(t)\int_\delta^t D_-\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds \end{aligned} \tag{2.11}$$

with $\delta\in[1,\infty)$.

We denote the three terms on the right-hand side of (2.11) by $u_{p_+}=H_+f$, $u_{p_0}=H_0f$, $u_{p_-}=H_-f$, respectively. De Hoog and Weiss (1980a) showed that u_{p_+} and u_{p_-} are in $C([1,\infty])$ if $D_-E^{-1}f$ and $D_+E^{-1}f$ are in $C([1,\infty])$ and that $J(D_++D_-)u_p(\infty)=-(D_++D_-)E^{-1}f(\infty)$ holds.

Lentini and Keller (1980a) showed that

$$\|u_{p_0}(t)\|=O(t^{-\varepsilon}) \quad \text{if} \quad \|D_0E^{-1}f\|=O(t^{-(\alpha+1)r-\varepsilon})$$

where $\varepsilon>0$ and r is the maximal dimension of the invariant subspaces of J associated with eigenvalues on the imaginary axis. Therefore the operator H operates on the space of all functions f, which fulfill

$$f\in C([1,\infty]) \quad \text{and} \quad D_0E^{-1}f(t)=F_0(t)t^{-(\alpha+1)r-\varepsilon} \tag{2.12}$$

with $F_0\in C_b([1,\infty))$, where $C_b([1,\infty))$ is the space of functions which are continuous on $[1,\infty)$ and bounded as $t\to\infty$.

Inserting the general solution of (2.5), (2.6) into (2.7) we get

$$\left[(B_1E+B_\infty E)\bar{D}_0+B_1E\exp\left(\frac{J}{\alpha+1}\right)D_-\right]\xi=\hat{\beta}-\left[B_1Eu_p(1)+B_\infty Eu_p(\infty)\right]. \tag{2.13}$$

Therefore we conclude the following:

THEOREM 2.1. *The problem* (2.1), (2.2), (2.3) *has a unique solution y for all f which fulfill* (2.12) *and $\hat{\beta}\in R^{\bar{r}_0+r_-}$, if and only if*

$$\operatorname{rank}\left[(B_1E+B_\infty E)\bar{D}_0+B_1\exp\left(\frac{J}{\alpha+1}\right)D_-\right]=\bar{r}_0+r_-, \tag{2.14}$$

where B_1 and B_∞ are $(\bar{r}_0+r_-)\times n$ matrices.

In this case y depends continuously (in the norm $\|y\|_{[1,\infty]}:=\max_{t\in[1,\infty]}\|y(t)\|$) on the data $\hat{\beta}$, $(D_++D_-)E^{-1}f$ and F_0. This follows directly from estimates given in the papers cited above. We see that (2.2) is an additional boundary condition at $t=\infty$ of the rank $r_++(r_0-\bar{r}_0)$.

Now we investigate the decay properties of u_p on the dependence of the decay properties of f.

THEOREM 2.2. *If f fulfills* (2.12) *then the following estimates hold for $t\geq\delta$:*

$$\|(H_+f)(t)\|\leq\text{const.}\|D_+E^{-1}f\|_{[t,\infty]}, \tag{2.15}$$

$$\|(H_0f)(t)\|\leq\text{const.}\,t^{-\varepsilon}\max_{s\geq t}\|s^{(\alpha+1)r+\varepsilon}D_0E^{-1}f(s)\|, \tag{2.16}$$

with $\varepsilon>0$.

For arbitrary $\gamma\geq0$

$$\|(H_-f)(t)\|\leq\text{const.}\,t^{-\gamma}\max_{\delta\leq s\leq t}\|s^{\gamma}D_-E^{-1}f(s)\|, \tag{2.17}$$

where all constants are independent of f and δ.

The first two estimates have been proven by Lentini and Keller (1980a). The third estimate follows from

$$\|u_{p_-}(t)\|<\text{const.}\max_{i=1(1)k}\left(\exp\left(-\frac{\lambda}{\alpha+1}t^{\alpha+1}\right)\int_\sigma^t\exp\left(\frac{\lambda}{\alpha+1}s^{\alpha+1}\right)(t^{\alpha+1}-s^{\alpha+1})^{i-1}s^{\alpha-\gamma}\,ds\right)$$
$$\max_{\delta\leq s\leq t}\|s^{\gamma}D_-E^{-1}f(s)\|,$$

where $-\lambda$ is the largest real part of eigenvalues of J with $\lambda>0$ and k is the maximal dimension of the associated Jordan-blocks.

By applying l'Hôpital's theorem to

$$g_i(t)=\frac{\int_\delta^t \exp\left(\frac{\lambda}{e^{\alpha+1}}s^{\alpha+1}\right)(t^{\alpha+1}-s^{\alpha+1})^{i-1}s^{\alpha-\gamma}\,ds}{\exp\left(\frac{\lambda}{e^{\alpha+1}}t^{\alpha+1}\right)t^{-\gamma}}$$

it is easy to conclude that $\lim_{t\to\infty} g_i(t)=\text{const.}$

In particular Theorem 2.2 tells us that inhomogeneities which converge to zero algebraically produce particular solutions which converge algebraically with exponent increased by $(\alpha+1)r$.

Now we want to investigate exponentially decreasing inhomogeneities.

If $\int_\delta^\infty D_-\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds$ exists we can substitute H_- by $\tilde{H}_-$ which is defined by

$$(\tilde{H}_-f)(t)=\phi(t)\int_\infty^t D_-\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds. \tag{2.18}$$

Now we prove the following theorem.

THEOREM 2.3. *Let J^- consist of Jordan blocks belonging to the same eigenvalue $-\lambda$ and let k be the dimension of the largest of these blocks. Furthermore, let $D_-E^{-1}f(t)=t^\beta\exp(-(\omega/\alpha+1)t^{\alpha+1})F_-(t)$ with $F_-\in C_b([1,\infty))$, $\beta\in R$ and $\omega>0$. Then for $t\geq\delta$:*

$$\|(H_-f)(t)\|\leq\text{const.}\,t^\beta\exp\left(-\frac{\omega}{\alpha+1}t^{\alpha+1}\right)\|F_-\|_{[\delta,t]} \tag{2.19}$$

if $\operatorname{Re}\lambda-\omega>0$,

$$\|(H_-f)(t)\|\leq\text{const.}\exp\left(-\frac{\omega}{\alpha+1}t^{\alpha+1}\right)t^{(\alpha+1)k+\beta}\ln t\|F_-\|_{[\delta,t]} \tag{2.20}$$

if $\operatorname{Re}\lambda-\omega=0$ *and* $\beta\geq-k(\alpha+1)$. *The factor* $\ln t$ *only appears if* $\beta=-k(\alpha+1)$.

$$\|(\tilde{H}_-f)(t)\|\leq\text{const.}\exp\left(-\frac{\omega}{\alpha+1}t^{\alpha+1}\right)t^{(\alpha+1)k+\beta}\ln t\|F_-\|_{[\delta,t]} \tag{2.21}$$

if $\operatorname{Re}\lambda-\omega=0$ *and* $\beta<-k(\alpha+1)$.

$$\|(\tilde{H}_-f)(t)\|\leq\text{const.}\exp\left(-\frac{\omega}{\alpha+1}t^{\alpha+1}\right)t^{\beta}\|F_-\|_{[t,\infty]} \tag{2.22}$$

if $\operatorname{Re}\lambda-\omega<0$. *All constants are independent of f and δ.*

The proofs are analogous to Theorem 2.2.

Theorem 2.3 implies that exponentially decaying inhomogeneities produce particular solutions which converge with the same exponential factor; however, the algebraic factor may change and a logarithmic factor may appear. If $\tilde{H}_-$ exists then it cuts off the terms of the particular solution which are already included in $\phi(t)D_-\xi$.

Assume now that J^- consists of more than one Jordan block with different eigenvalues and that $D_-E^{-1}f(t)$ has the form as in Theorem 2.3. Then H_- *and* $\tilde{H}_-$ may be used in order to gain a particular solution which decays as fast as possible according to the different cases of Theorem 2.3. Doing this, D_- has to be split up into the projections onto the direct sums of the invariant subspaces associated with different eigenvalues with negative real part and H_- and $\tilde{H}_-$ have to be applied to the resulting subsystems, respectively. We call the resulting operator $\bar{H}$. Its composition depends on the decay properties of f and on J^-.

3. Linear variable coefficient problems. Now we analyse

$$y'-t^{\alpha}A(t)y=t^{\alpha}f(t), \qquad \alpha\in N_0, \tag{3.1}$$

$$y\in C([1,\infty]), \tag{3.2}$$

$$B_1y(1)+B_{\infty}y(\infty)=\hat{\beta}. \tag{3.3}$$

The $n\times n$ matrix $A(t)$ fulfills

$$A\in C([1,\infty]), \qquad A(\infty)\neq 0, \tag{3.4}$$

$$A \text{ is analytic in } [\delta,\infty] \text{ for some } \delta\geq 1 \tag{3.5}$$

so that

$$A(t)=\sum_{i=0}^{\infty}A_it^{-i} \quad \text{for } t \text{ sufficiently large,} \tag{3.6a}$$

where

$$A_i=\frac{1}{i!}\lim_{x\to 0^+}\frac{d^i}{dx^i}A\left(\frac{1}{x}\right). \tag{3.6b}$$

Let J_0 be the Jordan canonical form of A_0 obtained by the transformation

$$A_0=EJ_0E^{-1} \tag{3.7}$$

and let the J_i's be defined by

$$A_i=EJ_iE^{-1}. \tag{3.8}$$

The matrices J_i are the coefficients of the series

$$J(t)=E^{-1}A(t)E=\sum_{i=0}^{\infty}J_it^{-i} \quad \text{for } t\to\infty. \tag{3.9}$$

We set

$$J_0=\operatorname{diag}(J_0^+,J_0^0,J_0^-), \quad \dim(J_0^+)=r_+, \quad \dim(J_0^0)=r_0, \quad \dim(J_0^-)=r_-, \tag{3.10}$$

where J_0^+, J_0^0, J_0^- have only eigenvalues with positive, zero and negative real parts, respectively.

Again we substitute

$$u=E^{-1}y \tag{3.11}$$

and get the problem

$$u'-t^{\alpha}J(t)u=t^{\alpha}E^{-1}f(t), \qquad 1\leq t<\infty, \tag{3.12}$$

$$u\in C([1,\infty]). \tag{3.13}$$

For the following we need the definition of an asymptotic series. A function $P(t)$ is said to be represented asymptotically by a formal series (in the Poincaré sense)

$$P(t)\sim\sum_{i=0}^{\infty}P_it^{-i}, \qquad t\to\infty, \tag{3.14}$$

if

$$t^m\left[P(t)-\sum_{i=0}^{m}P_it^{-i}\right]\to 0 \quad \text{for } t\to\infty \text{ and } m\geq 0. \tag{3.15}$$

Therefore,

$$P(t)=\sum_{i=0}^{m} P_i t^{-i}+O(t^{-m-1}) \quad \text{for } t\to\infty \text{ and } m\geq 0 \tag{3.16}$$

holds. To get more information on asymptotic series see Wasow (1965).

Coddington and Levinson (1955) and Wasow (1965) prove an asymptotic representation of the fundamental matrix of the homogeneous system (3.12) which we state in the following theorem.

THEOREM 3.1. *Under the given assumptions on $J(t)$ there is a fundamental matrix which has the form*

$$\phi(t)=P(t)t^{D}e^{Q(t)},$$

where $Q(t)$ is a diagonal matrix:

$$Q(t)=\operatorname{diag}(J_0)\frac{t^{\alpha+1}}{\alpha+1}+Q_1\frac{t^{\alpha+1-(1/p)}}{p(\alpha+1)-1}+Q_2\frac{t^{\alpha+1-(2/p)}}{p(\alpha+1)-2}$$
$$+\cdots+Q_{p(\alpha+1)-2}\frac{t^{2/p}}{2}+Q_{p(\alpha+1)-1}t^{1/p}$$

with some $p\in N$, D is a constant matrix in Jordan canonical form and

$$P(t)=P_1(t)P_2(t),$$

where

$$P_1(t)\sim I+\sum_{i=1}^{\infty} P_{1i}t^{-i}, \qquad t\to\infty$$

and

$$P_2(t)\sim\sum_{i=0}^{\infty} P_{2i}t^{-i/p}, \qquad t\to\infty.$$

To every Jordan block of D correspond equal (diagonal) entries of $Q(t)$. Therefore t^D and $e^{Q(t)}$ commute. Moreover, the block structure of D is a subdivision of that blocking of J_0 which is obtained by gathering all Jordan blocks of J_0 belonging to the same eigenvalue. Also $P_2(t)$ has a block structure which is identical to the above mentioned blocking of J_0.

In the case of distinct eigenvalues of $A(\infty)$, $p=1$ holds, D is diagonal and $P(\infty)=I$.

The proof of this asymptotic expansion for $\phi(t)$ given by Wasow (1965) is constructive and therefore contains an algorithm for the calculation of P, D and Q. We present an outline of the construction of $\phi(t)$ since it will be needed for setting up particular solutions. We assume that J_0 has the different eigenvalues $\lambda_1,\cdots,\lambda_k$ and the block diagonal form

$$J_0=\operatorname{diag}\left(J_0^{(i)},\cdots,J_0^{(k)}\right), \qquad \dim\left(J_0^{(i)}\right)=r_i; \tag{3.17}$$

$J_0^{(i)}$ has the only eigenvalue λ_i. Then the following algorithm results.

Step 1. Substitute

$$u=P_1(t)\begin{pmatrix} u_{(1)} \\ \vdots \\ u_{(k)} \end{pmatrix}$$

and determine

$$P_1(t)\sim I+\sum_{i=1}^{\infty} P_{li}t^{-i}, \qquad t\to\infty$$

such that the resulting system (with

$$\begin{pmatrix} u_{(1)} \\ \vdots \\ u_{(k)} \end{pmatrix}$$

as dependent variable) splits up into k separate subsystems of the form

$$u'_{(i)}=t^{\alpha}J_{(i)}(t)u_{(i)}, \tag{3.18}$$

with

$$J_{(i)}(t)\sim J_0^{(i)}+\sum_{j=1}^{\infty} J_j^{(i)}t^{-j}. \tag{3.19}$$

Step 2. We substitute

$$u_{(i)}=v_{(i)}\exp\left(\lambda_i\frac{t^{\alpha+1}}{\alpha+1}\right) \tag{3.20}$$

and get

$$v'_{(i)}=t^{\alpha}\left(J_{(i)}(t)-\lambda_i I_{r_i}\right)v_{(i)} \quad \text{for } i=1(1)k. \tag{3.21}$$

The leading matrices of the systems (3.21) are now $J_0^{(i)}-\lambda_i I_{r_i}$ having the only eigenvalue 0.

Step 3. We apply so-called shearing transformations

$$v_{(i)}=S_{(i)}(t)w_{(i)}, \qquad i=1(1)k, \tag{3.22}$$

where

$$S_{(i)}(t)=\operatorname{diag}\left(1,t^{-g_i},t^{-2g_i},\cdots,t^{-(r_i-1)g_i}\right), \tag{3.23}$$

with $g_i\geq 0$ (and rational) to the systems (3.21). The g_i's are chosen such that the leading matrices of the resulting systems, which have $w_{(i)}$ as dependent variables, have more than one different eigenvalue or, if not possible for a certain i, such that the rank of this new ith system is smaller than $\alpha+1$ or that this system splits up into separate subsystems.

Wasow (1965) showed that it is always possible to achieve one of these simplifications. In order to get systems where only integral powers of the independent variable occur we substitute

$$x_i=p_i^{1/(g_i-\alpha-1)}t^{1/p_i} \quad \text{for } i=1(1)k, \tag{3.24}$$

where p_i is the smallest integer so that $g_i p_i$ is an integer. Then we get systems of the form

$$w'_{(i)}(x_i)=x_i^{h_i}C_{(i)}(x_i)w_{(i)}(x_i), \qquad i=1(1)k \tag{3.25}$$

where

$$h_i=(\alpha+1-g_i)p_i-1, \tag{3.26}$$

$$C_{(i)}(x_i)\sim\sum_{j=0}^{\infty} C_j^{(i)}x_i^{-j}. \tag{3.27}$$

Step 4a. If $C_0^{(i)}$ has only 0 as eigenvalue then we have reduced the rank of the system or it splits up into separate subsystems of lower order. Applying more shearing transformations to (3.25) we end up with a system whose leading matrix has either more than one different eigenvalue or has rank equal to 0. In the second case we have a system with a singularity of the first kind for which the fundamental matrix $\phi_{(i)}(t)$ has the form

$$\phi_{(i)}(t)=P_{(i)}(t)t^{D_{(i)}}, \qquad P_{(i)}(t)=\sum_{j=0}^{\infty} P_j^{(i)}t^{-j} \tag{3.28}$$

and $D_{(i)}$ is a constant matrix in Jordan canonical form (see Wasow (1965)).

Step 4b. Now we assume that $C_0^{(i)}$ has at least two different eigenvalues. Then we transform $C_0^{(i)}$ to its Jordan canonical form $\tilde{C}_0^{(i)}$,

$$C_0^{(i)}=E_{(i)}\tilde{C}_0^{(i)}E_{(i)}^{-1}, \tag{3.29}$$

and substitute

$$w_{(i)}(x_i)=E_{(i)}z_{(i)}(x_i), \tag{3.30}$$

getting a system whose leading matrix is $\tilde{C}_0^{(i)}$, which means, in Jordan form,

$$z'_{(i)}(x_i)=x_i^{h_i}C_{(i)}(x_i)z_{(i)}(x_i), \qquad \tilde{C}_{(i)}(\infty)=\tilde{C}_0^{(i)}. \tag{3.31}$$

Step 5. We apply the transformation given in Step 1 to the system (3.31) in order to get separate subsystems of lower order whose leading matrices have the different eigenvalues $\mu_{(i)j}$. By the means of Step 2 we normalize these systems so that their leading matrices have only the eigenvalue 0. These transformations split off the factors

$$\exp\left(\mu_{(i)j}\frac{x_i^{h_i+1}}{h_i+1}\right) \quad \text{for } i=1(1)k. \tag{3.32}$$

Resubstituting (3.24) and using (3.26) we notice that the argument in (3.32) is of order $t^{\alpha+1-g_i}$, that means of order lower than $t^{\alpha+1}$ which is the order of the argument of the first exponential factor because if $g_i=0$ the system remained unchanged.

Applying another set of shearing transformations as in Step 3 we arrive at Step 4a or Step 4b.

Step 6. A finite chain of all the described transformations in Step 1 to Step 5 result in a set of one-dimensional systems and systems with a singularity of the first kind. Setting $p=p^{-m}$, where $\bar{p}$ is the smallest common multiple of all the p's used in the sequence of shearing transformations and m is the number of these transformations which split the system into a set of systems described above, we get the formula for the fundamental matrix $\phi(t)$ given in Theorem 3.1 by taking into account (3.28). Moreover we get

$$P_2(t)=\left(\prod_{l=1}^{m} S_l(t)E_lP_{3(l)}(t)\right)P_4(t), \qquad P_{3(l)}(t)=I+\sum_{j=1}^{\infty} P_{3j}^{(l)}t^{-j/p} \tag{3.33}$$

and the $S_l(t)$ are composed of submatrices $S_{(lj)}(t)$ defined in (3.23). The E_l's are nonsingular and $P_4(t)$ is derived by solving the systems with singularities of the first kind using (3.28).

We define

$$P_3(t)=\prod_{l=1}^{m} S_l(t)E_lP_{3(l)}(t). \tag{3.34}$$

An estimate of $P_3^{-1}(t)$ can be obtained as follows.

Let D_i be the projection onto the direct sum of invariant subspaces associated with the eigenvalue λ_i of J_0. Then

$$\|P_3^{-1}(t)D_i\| \leq \text{const.}\, t^{[(r_i-1)g_{i1}+(g_{i2}/p_{i1})+(g_{i3}/p_{i1}p_{i2})+\cdots+(g_{im}/p_{i1}\cdots p_{i,m-1})]} \tag{3.35}$$

holds. The sum in the exponent of t is derived by estimating $S_l^{-1}(t)$ and by taking into account the block structures of the E_l^{-1} and $P_{3(l)}^{-1}(t)$. p_{il} and g_{il} are as in (3.23) and represent that sequence of shearings starting off from the ith r_i-dimensional subsystem and giving the largest exponent in (3.35).

For this sequence we calculate the ranks of the corresponding sequence of subsystems as in (3.26)

$$\begin{aligned} &h_{i0}+1=\alpha+1,\\ &h_{i1}+1=p_{i1}(h_{i0}+1-g_{i1}),\\ &\vdots\\ &h_{im}+1=p_{im}(h_{i,m-1}+1-g_{im}). \end{aligned} \tag{3.36}$$

By assumption $h_{im}+1\geq 0$ holds and so we get

$$\alpha+1>g_{i1}+\frac{g_{i2}}{p_{i1}}+\cdots+\frac{g_{im}}{p_{i1},\cdots,p_{i,m-1}} \tag{3.37a}$$

and therefore the estimate

$$\|P_3^{-1}(t)D_i\| \leq \text{const.}\, t^{(r_i-1)(\alpha+1)} \tag{3.37b}$$

holds.

The basic solutions φ_i with $\phi(t)=(\varphi_1(t),\cdots,\varphi_1(t))$ fulfill

$$\|\varphi_i(t)\| \leq p_i(t)e^{q_i(t)}t^{d_i}(\ln t)^{j_i}, \qquad p_i\in C([1,\infty]). \tag{3.38}$$

Eigenvalues of J_0 with positive real part produce exponentially increasing basic solutions; eigenvalues with negative real part produce exponentially decreasing basic solutions. Imaginary eigenvalues of J_0 can produce exponentially and algebraically increasing and decreasing, constant and oscillating and logarithmically increasing solutions. The asymptotic behaviour of a particular basic solution φ_i can be determined by knowing D and $P_0,\cdots,P_{m_i}$ where m_i is sufficiently large. Therefore the solution of the homogeneous problem (3.12), (3.13) is

$$u_h(t)=\phi(t)(\tilde{D}_0+D_-)\xi, \qquad \xi\in C^n, \tag{3.39}$$

where D_- is as in §2 and the diagonal projection $\tilde{D}_0$ sorts out the solution columns of $\phi(t)$ which are in $C([1,\infty])$ and which are produced by eigenvalues with zero real part. We define $\tilde{r}_0$ as the number of nonzero columns of $\tilde{D}_0$.

Now we construct a particular solution $u_p=Hf$ of the problem

$$u_p'=t^\alpha J(t)u_p+t^\alpha E^{-1}f(t), \qquad 1\leq t<\infty, \quad f\in C([1,\infty]), \tag{3.40}$$

$$u_p\in C([1,\infty]). \tag{3.41}$$

We substitute

$$u_p(t)=P_1(t)v_p(t), \qquad v_p(t)=\begin{bmatrix} v_{p_+}(t)\\ v_{p_0}(t)\\ v_{p_-}(t)\end{bmatrix} \tag{3.42}$$

and define

$$(3.43)\quad u_{p_+}(t)=P_1(t)\begin{bmatrix} v_{p_+}(t)\\ 0\\ 0\end{bmatrix},\quad u_{p_0}(t)=P_1(t)\begin{bmatrix} 0\\ v_{p_0}(t)\\ 0\end{bmatrix},\quad u_{p_-}(t)=P_1(t)\begin{bmatrix} 0\\ 0\\ v_{p_-}(t)\end{bmatrix}.$$

We get the separated problems

$$\begin{bmatrix} v_{p_+}\\ v_{p_0}\\ v_{p_-}\end{bmatrix}=t^{\alpha}\big(\operatorname{diag}\big(J_0^+ +J^+(t),J_0^0+J^0(t),J_0^- +J^-(t)\big)\big)\begin{bmatrix} v_{p_+}\\ v_{p_0}\\ v_{p_-}\end{bmatrix}+t^{\alpha}P_1^{-1}(t)E^{-1}f(t),$$

where $J^+(t)$, $J^0(t)$, $J^-(t)$ have an asymptotic power series expansion in t^{-1} without a constant term. We define the block components v_{p_+}, v_{p_-}, as in de Hoog and Weiss (1980a,b), as solutions of the operator equations

$$(3.44a)\qquad v_{p_+}=\hat{H}_+J_0^+v_{p_+}+\hat{H}_+\big(P_1^{-1}E^{-1}f\big)_+,$$

$$(3.44b)\qquad v_{p_-}=\hat{H}_-J_0^-v_{p_-}+\hat{H}_-\big(P_1^{-1}E^{-1}f\big)_-,$$

where $(P_1^{-1}E^{-1}f)_+$ and $(P_1^{-1}E^{-1}f)_-$ are the first r_+ and the last r_- components respectively of $P_1E^{-1}f$ and $\hat{H}_+$, $\hat{H}_-$ are the operators defined similar to (2.11):

$$(3.44c)\qquad \big(\hat{H}_+g_+\big)(t)=\int_{\infty}^{t}\exp\Big(\frac{J_0^+}{\alpha+1}(t^{\alpha+1}-s^{\alpha+1})\Big)s^{\alpha}g_+(s)\,ds,$$

$$(3.44d)\qquad \big(\hat{H}_-g_-\big)(t)=\int_{\delta}^{t}\exp\Big(\frac{J_0^-}{\alpha+1}(t^{\alpha+1}-s^{\alpha+1})\Big)s^{\alpha}g_-(s)\,ds$$

for $g_{\pm}\in C([1,\infty])$.

From (3.44) we derive

$$(3.45a)\qquad v_{p_+}=\big(I-\hat{H}_+J^+\big)^{-1}\hat{H}_+\big(P_1^{-1}E^{-1}f\big)_+\in C([\delta,\infty]),$$

$$(3.45b)\qquad v_{p_-}=\big(I-\hat{H}_-J^-\big)^{-1}\hat{H}_-\big(P_1^{-1}E^{-1}f\big)_-\in C([\delta,\infty])$$

with δ sufficiently large. The proof of the invertibility of $(I-H_{+,-}J^{+,-})$ is given in de Hoog and Weiss (1980a,b). Again we get

(3.46)

$$\text{(a)}\quad v_{p_+}(\infty)=-(J_0^+)^{-1}\big(E^{-1}f(\infty)\big)_+,\qquad \text{(b)}\quad v_{p_-}(\infty)=-(J_0^-)^{-1}\big(E^{-1}f(\infty)\big)_-,$$

because $J^+(\infty)=0$ and $J^-(\infty)=0$.

Now we assume that for some $\varepsilon>0$,

$$(3.47)\qquad D_0P_1^{-1}(t)E^{-1}f(t)=F_0(t)t^{-\bar{r}(\alpha+1)-\varepsilon},\qquad F_0\in C_b([1,\infty)),$$

holds, where $\bar{r}$ is the largest algebraic multiplicity of the eigenvalues of J_0 with real part zero. So $\bar{r}$ is defined differently to r in §2.

The system

$$(3.48)\qquad v_{p_0}'=t^{\alpha}\big(J_0^0+J^0(t)\big)v_{p_0}$$

is composed of separate systems, each of them associated with one imaginary eigenvalue of J_0 and $\bar{r}$ is the maximal dimension of these subsystems.

We take one of these (inhomogeneous) subsystems

$$(3.49)\qquad v_{p_{O(i)}}'=t^{\alpha}J_{(i)}(t)v_{p_{O(i)}}+t^{\alpha}\big(P_1^{-1}(t)E^{-1}f(t)\big)_{O(i)},$$

where $\hat{f}_{(i)}(t) := (P^{-1}(t)E^{-1}f(t))_{o(i)}$ consists of the corresponding components of the inhomogeneity $P^{-1}(t)E^{-1}f(t)$. The leading matrix of $J_{(i)}(t)$ has the only eigenvalue λ_i with $\mathrm{Re}\lambda_i = 0$.

Now we apply the transformations

$$v_{p_{O(i)}} = \exp\left(\frac{\lambda_i t^{\alpha+1}}{\alpha+1}\right) S_{(i)}(t) w_{p_{O(i)}}, \qquad x_i = c(i)t^{1/p_i}, \tag{3.50}$$

$$w_{p_{O(i)}}(x_i) = E_{(i)} z_{p_{O(i)}}(x_i) \tag{3.51}$$

as defined in (3.20), (3.23), (3.24) and (3.30) and get

$$z'_{p_{O(i)}}(x_i) = x_i^{h_i}\tilde{C}_{(i)}(x_i) z_{p_{O(i)}}(x_i) + x_i^{h_i}\hat{g}_{(i)}(x_i), \tag{3.52}$$

where

(3.53)

$$\hat{g}_{(i)}(x_i) = \left(\frac{x_i}{c(i)}\right)^{p_i g_i} E_{(i)}^{-1} S_{(i)}^{-1}\left(\left(\frac{x_i}{c(i)}\right)^{p_i}\right) \exp\left(-\frac{\lambda_i}{\alpha+1}\left(\frac{x_i}{c(i)}\right)^{p_i(\alpha+1)}\right) \hat{f}_{(i)}\left(\left(\frac{x_i}{c(i)}\right)^{p_i}\right).$$

From (3.37b), (3.47) we derive:

$$\|\hat{g}_{(i)}(x_i)\| \leq \text{const.}\, x_i^{-p_i\varepsilon - \bar{r}p_i(\alpha+1-g_i)} \|F_0\|_{[1,\infty]}. \tag{3.54}$$

If the leading matrix of the system (3.52) has at least one eigenvalue different from zero then the separating transformation

$$z_{p_{O(i)}} = P_{3(i)}(x_i) z^{(1)}_{p_{O(i)}}, \qquad P_{3(i)}(\infty) = I,$$

can be applied. Those resulting subsystems, whose leading matrices have eigenvalues with real part different from zero can now be solved by the means of (3.44), (3.45) because the new inhomogeneity has the form

$$\hat{g}^1_{(i)}(x_i) = P^{-1}_{3(i)}(x_i)\hat{g}_{(i)}(x_i) \in C([1,\infty]). \tag{3.55}$$

For all other subsystems this sequence of substitutions is repeated as long as we arrive at systems whose leading matrices have eigenvalues with real part different from zero or one-dimensional systems or we arrive at systems with a singularity of rank zero. In the second and third cases only eigenvalues with real part zero have been split off, therefore the inhomogeneities do not contain exponentially increasing or decreasing factors (see (3.53)). For every system

$$z' = x^h C(x) z + x^h g_0(x) \tag{3.56a}$$

that occurs in this sequence of transformations we get

$$\|g_0(x)\| \leq \text{const.}\, x^{-\tilde{\varepsilon} - \bar{r}(h+1)} \|F_0\|_{[1,\infty]}, \tag{3.56b}$$

where $\tilde{\varepsilon} > 0$ is such that after resubstitution to t as independent variable $x(t)^{-\tilde{\varepsilon}} \leq \text{const.}\, t^{-\varepsilon}$ holds. (3.56b) follows from (3.37a) and (3.47).

A particular solution for one-dimensional systems can be found easily, so we just have to treat systems with a singularity of the first kind.

We want to solve

$$z'_p = \left(\frac{1}{x}B + \frac{1}{x}\tilde{B}(x)\right) z_p + \frac{1}{x} g_0(x), \qquad \tilde{B}(x) = \bar{B}(x)\frac{1}{x}, \tag{3.57}$$

$$z_p(x) = Z_p(x) x^{-\varepsilon} (\ln x)^j, \quad Z_p \in C_b([1,\infty)), \quad j \in N_0, \tag{3.58}$$

where B is a constant matrix in Jordan form, $\bar{B} \in C([1,\infty])$ and g_0 fulfills (3.56).

Let

$$B=\operatorname{diag}(B_1,\cdots,B_q),\qquad B_i=\begin{bmatrix} & 1 & \\ & \ddots & \ddots \\ & & 1 \\ b_i & & \end{bmatrix},\quad i=1(1)q, \tag{3.59}$$

and let D^i be the projection onto the invariant subspace associated with b_i.

Then we define

$$z_p=G\tilde{B}z_p+Gg_0,\qquad G=G_1+\cdots+G_q, \tag{3.60}$$

and

$$(G_ig_0)(x)=\begin{cases} x^B\int_\delta^x D^i s^{-B-I}g_0(s)\,ds, & b_i\le-\varepsilon,\\ x^B\int_x^\infty D^i s^{-B-I}g_0(s)\,ds, & b_i>-\varepsilon. \end{cases} \tag{3.61}$$

It is easily checked that for $x\ge\delta$

$$\|(Gg_0)(x)\|\le\text{const.}(\ln x)^j\max\Big(\max_{s\in[x,\infty]}\|g_0(s)\|,\ x^{-\varepsilon}\max_{s\in[\delta,x]}\|s^\varepsilon g_0(s)\|\Big) \tag{3.62}$$

holds where $j=\max_i(\dim(B_i))$. Therefore we get from (3.60)

$$z_p=(I-G\tilde{B})^{-1}Gg_0, \tag{3.63}$$

z_p fulfills (3.58). If $g_0(x)=O(x^{-\varepsilon}(\ln x)^i)$ then the right-hand side of (3.62) has an additional factor $(\ln x)^i$.

After having performed all necessary resubstitution we get a solution $v_{p_0}(t)$ fulfilling $v_{p_0}(\infty)=0$. Defining $\tilde{D}_{00}$ as the projection onto the columns of $\phi(t)$ which tend to a nonzero limit as $t\to\infty$ we get the following theorem.

THEOREM 3.2. *The problem* (3.1), (3.2), (3.3) *under the assumptions* (3.4), (3.5) *has a unique solution y for all $f\in C([1,\infty])$ fulfilling* (3.47) *and for all $\hat{\beta}\in R^{\tilde{r}_0+r_-}$ if and only if*

$$\operatorname{rank}\big[B_1E\phi(1)(\tilde{D}_0+D_-)+B_\infty E\tilde{D}_{00}\big]=\tilde{r}_0+r_-$$

where B_1 and B_∞ are $(\tilde{r}_0+r_-)\times n$ matrices. This solution y depends continuously on $\hat{\beta}$, $(D_++D_-)E^{-1}P_1^{-1}f$ and F_0 which is defined in (3.47).

The proof is complete if the continuity statement is proven.

THEOREM 3.3. *If $f\in C([1,\infty])$ and* (3.47) *holds then the following estimates hold for $t\ge\delta$:*

$$\|(H_+f)(t)\|\le\text{const.}\,\|D_+P_1^{-1}E^{-1}f\|_{[t,\infty]}. \tag{3.64}$$

For arbitrary $\gamma\ge0$, $\varepsilon\ge0$

$$\|(H_-f)(t)\|\le\text{const.}\,t^{-\gamma}\max_{s\in[\delta,t]}\|s^\gamma D_-P_1^{-1}(s)E^{-1}f(s)\|\quad \text{for } \gamma>0. \tag{3.65}$$

(3.66)

$$\|(H_0f)(t)\|\le\text{const.}(\ln t)^{j_0}\max\Big(t^{-\varepsilon}\max_{s\in[\delta,t]}\|s^{\varepsilon+(\alpha+1)\bar{r}}D_0P_1^{-1}(s)E^{-1}f(s)\|,\ \max_{s\ge t}\|s^{(\alpha+1)\bar{r}}D_0P_1^{-1}(s)E^{-1}f(s)\|\Big),$$

where $j_0=\max_{i_k}(\dim(D_{i_k}))$ and D_{i_k} are the Jordan blocks of the matrix D in Theorem 3.1 *for which the corresponding polynomials $q_{i_k}(t)$ fulfill $\operatorname{Re}q_{i_k}(t)\equiv0$. All constants are independent of δ and f.*

Proof. From (3.45) we conclude

$$v_{p_+}(t)=\sum_{i=0}^{\infty}\left(\left(\hat{H}_+J^+\right)^i\hat{H}_+\left(P_1^{-1}E^{-1}f\right)_+\right)(t), \tag{3.67a}$$

$$v_{p_-}(t)=\sum_{i=0}^{\infty}\left(\left(\hat{H}_-J^-\right)^i\hat{H}_-\left(P_1^{-1}E^{-1}f\right)_-\right)(t) \tag{3.67b}$$

for δ sufficiently large.

$\hat{H}_+$ (resp. $\hat{H}_-$) fulfill the estimates (2.15) (resp. (2.17)) and therefore we get:

$$\left\|v_{p_+}(t)\right\|\leq \text{const.}\max_{s\in[t,\infty]}\left\|D_+P_1^{-1}(s)E^{-1}f(s)\right\|\sum_{i=0}^{\infty}\left(\frac{1}{2}\right)^i, \tag{3.68a}$$

$$\left\|v_{p_-}(t)\right\|\leq \text{const.}\,t^{-\gamma}\max_{s\in[\delta,t]}\left\|s^{\gamma}D_-P_1^{-1}(s)E^{-1}f(s)\right\|\sum_{i=0}^{\infty}\left(\frac{1}{2}\right)^i \tag{3.68b}$$

if δ is so large that $\|J^+\|_{[\delta,\infty]}\leq\frac{1}{2}$ and $\|J^-\|_{[\delta,\infty]}\leq\frac{1}{2}$. The estimates (3.64), (3.65) follow by using (3.39) and (3.66) follows from the derivation of v_{p_0} and from (3.62), (3.63).

Theorem 3.3 implies that an inhomogeneity f fulfilling

$$f(t)=t^{-(\alpha+1)\bar{r}-\varepsilon}(\ln t)^l F(t),\qquad F\in C_b([1,\infty)), \tag{3.69}$$

produces a particular solution Hf for which the estimate

$$\left\|(Hf)(t)\right\|\leq\text{const.}\,t^{-\varepsilon}(\ln t)^{j_0+l}\|F\|_{[\delta,\infty]} \tag{3.70}$$

holds.

Now we take inhomogeneities of the form

$$f(t)=e^{p(t)}t^{\beta}(\ln t)^l F(t),\qquad F\in C_b([1,\infty)), \tag{3.71}$$

where f fulfills (3.47). $p(t)$ is a polynomial in $t^{1/p}$.

We can construct a particular solution $\overline{H}f=H_+f+\overline{H}_0f+\overline{H}_-f$ fulfilling

$$\left\|\left(\overline{H}_0f(t)\right)\right\|\leq\text{const.}\,e^{p(t)}t^{\beta+\bar{r}(\alpha+1)}(\ln t)^{j_0+l}\|F\|_{[\delta,\infty]}, \tag{3.72a}$$

$$\left\|\left(\overline{H}_-f\right)(t)\right\|\leq\text{const.}\,e^{p(t)}t^{\beta+\bar{k}(\alpha+1)}(\ln t)^{j_-+l}\|F\|_{[\delta,\infty]}, \tag{3.72b}$$

where $\bar{k}$ is the maximal algebraic multiplicity of the eigenvalues of J_0 with negative real part and j_- is the maximal dimension of Jordan blocks in DD_-. The construction of $\overline{H}$ is similar to the construction of H.

Now we assume that the matrix $A(t)$ of the system (3.1) and $F(\tau)=A(1/\tau)$ fulfill

$$F\in C^{(\alpha+1)\bar{l}+1}\left(\left[0,\frac{1}{\delta}\right]\right),\qquad \delta\geq 1, \tag{3.73a}$$

$$A\in C([1,\infty]) \tag{3.73b}$$

instead of (3.5), where $\bar{l}=\max(\bar{r},\bar{k})$ is defined for the Jordan form of $A(\infty)$. Therefore,

$$A(t)=A_0+t^{-1}A_1+\cdots+t^{-(\alpha+1)\bar{l}}A_{(\alpha+1)\bar{l}}+\tilde{A}(t),\qquad \tilde{A}(t)=\bar{A}(t)t^{-(\alpha+1)\bar{l}-1-\beta} \tag{3.74}$$

holds, where $\bar{A}\in C_b([1,\infty))$, $\beta\geq 0$.

The system (3.1) can now be written as:

$$y'=t^{\alpha}\left(A_0+\cdots+t^{-(\alpha+1)\bar{l}}A_{(\alpha+1)\bar{l}}\right)y+t^{\alpha}\left(\tilde{A}(t)y(t)+f(t)\right). \tag{3.75}$$

The homogeneous problem (3.75) has the general solution

$$y_h = E\phi(\tilde{D}_0 + D_-)\xi + E\overline{H}_1\tilde{A}y_h, \qquad \xi \in C^n, \tag{3.76}$$

where $E\phi$ is the fundamental matrix of the unperturbed problem

$$\tilde{y}_h = t^{\alpha}\left(\sum_{i=0}^{(\alpha+1)\bar{l}} A_i t^{-i}\right)\tilde{y}_h.$$

Now let

$$\|\phi(t)(\tilde{D}_0 + D_-)\| = p(t)t^d e^{q(t)}(\ln t)^{\max(j_0, j_-)} = p(t)\sigma_h(t), \qquad p \in C([1,\infty]). \tag{3.77}$$

$\overline{H}_1$ is composed so that inhomogeneities which decay as $t^{-(\alpha+1)\bar{l}-1-\beta}\sigma_h(t)$ produce fast decaying particular solutions with regard to (3.69)–(3.72).

From (3.76) we get the equation

$$(I - E\overline{H}_1\tilde{A})y_h = E\phi(\tilde{D}_0 + D_-)\xi \tag{3.78}$$

for which we take as basic Banach-space

$$\left(A_{\sigma_h,\delta} = \{u \mid u(t) = U(t)\sigma_h(t), U \in C_b([\delta,\infty))\}, \qquad \|u\|_{\sigma_h,\delta} = \|U\|_{[\delta,\infty]}\right). \tag{3.79}$$

If $\sigma_j(t) \equiv 1$ we set $A_{\sigma_h,\delta} = C([\delta,\infty])$.

We get the estimate

$$\|\overline{H}_1\tilde{A}\|_{\sigma_h,\delta} = \max_{\|y_h\|_{\sigma_h,\delta}\le 1} \|\overline{H}_1\tilde{A}y_h\|_{\sigma_h,\delta} \le \text{const.}\,\delta^{-1-\beta}(\ln\delta)^{\max(j_0,j_-)} < \frac{1}{2\|E\|} \tag{3.80}$$

if δ is sufficiently large. Therefore $(I - E\overline{H}_1\tilde{A})^{-1}$ exists as an operator on $A_{\sigma_h,\delta}$ and, for $\xi \in C^n$,

$$y_h = (I - E\overline{H}_1\tilde{A})^{-1}E\phi(\tilde{D}_0 + D_-)\xi = \psi_-^0(\tilde{D}_0 + D_-)\xi \in A_{\sigma_h,\delta}. \tag{3.81}$$

As particular solution y_p of (3.43) we set

$$y_p = E\overline{H}_3\tilde{A}y_p + E\overline{H}_2 f. \tag{3.82}$$

The inhomogeneity f fulfills

$$\|f(t)\| = O(t^{\bar{d}}e^{\bar{q}(t)}) \quad \text{and} \quad \sigma_p(t) = t^{(\alpha+1)\bar{l}+\bar{d}}(\ln t)^{\max(j_0,j_-)}e^{\bar{q}(t)} \to 0 \tag{3.83}$$

and $\overline{H}_2$ is composed so that $(\overline{H}_2 f)(t)$ decays as fast as possible with regard to (3.69)–(3.72). Then

$$\|(\overline{H}_2 f)(t)\| = O(\sigma_p(t)); \tag{3.84}$$

$\overline{H}_3$ is composed to make particular solutions belonging to inhomogeneities which decay as $t^{-(\alpha+1)\bar{l}-1-\beta}\sigma_p(t)$ decrease as fast as possible. As basic space we now take $A_{\sigma_p,\delta}$ and conclude the invertibility of $(I - E\overline{H}_3\tilde{A})$ on $A_{\sigma_p,\delta}$, with δ sufficiently large and get

$$y_p = \psi(f) = (I - EH_3\tilde{A})^{-1}E\overline{H}_2 f \in A_{\sigma_p,\delta}. \tag{3.85}$$

Obviously,

$$y(t) = y_h(t) + y_p(t) \in C([\delta,\infty]) \tag{3.86}$$

holds. By substituting H, which is defined by (3.70) for $\overline{H}_1$ in (3.76), it is easily shown that the solution manifold y_h (with the parameters $(\tilde{D}_0 + D_-)\xi$) is unique in $C([\delta,\infty])$,

because $A_{\sigma_h,\delta} \subset C([\delta,\infty])$ and because the solution space is $\tilde{r}_0+r_-$ dimensional. Therefore y defined in (3.86) is unique in $C([\delta,\infty])$ (as manifold).

In order to get the solution in $C([1,\infty])$ we solve the "regular" problem:

$$y'=t^{\alpha}A(t)y+t^{\alpha}f(t), \qquad 1\le t\le\delta, \tag{3.87}$$

$$y(\delta)=y_h(\delta)+y_p(\delta). \tag{3.88}$$

From (3.76) and (3.81) we get, using the expansion

$$(I-G)^{-1}=\sum_{i=0}^{\infty} G^i \quad \text{for } \|G\|<1,$$

the following estimates which hold for $t\ge\delta$:

$$\left\|\psi_-^0(t)-E\phi(t)\left(\tilde{D}_0+D_-\right)\right\|\le \text{const.}\, t^{-1-\beta}(\ln t)^{\max(j_0,j_-)}\sigma_h(t) \tag{3.89}$$

and

$$\left\|(\phi(f))(t)-E\left(\bar{H}_2 f\right)(t)\right\|\le \text{const.}\, t^{-1-\beta}(\ln t)^{\max(j_0,j_-)}\sigma_p(t). \tag{3.90}$$

Theorem 3.2 remains valid if $E\phi(1)(\tilde{D}_0+D_-)$ is substituted by $\psi_-^0(1)$ (where ψ_-^0 has been continued to $[1,\infty]$).

Moreover it is important to consider problems where the matrix $\tilde{A}(t)$ defined in (3.74) decays exponentially

$$\tilde{A}(t)=\bar{A}(t)t^{\gamma}e^{q(t)}, \quad \bar{A}\in C_b([1,\infty)), \quad q(t)\to-\infty. \tag{3.91}$$

Using the same methods as in the case of algebraic decay we get:

$$\left\|\psi_-^0(t)-E\phi(t)\left(\tilde{D}_0+D_-\right)\right\|\le \text{const.}\, t^{\gamma+(\alpha+1)\bar{l}}(\ln t)^{\max(j_0,j_-)}\sigma_h(t)e^{q(t)}, \tag{3.92}$$

$$\left\|\left(\psi_p(f)\right)(t)-E\left(\bar{H}f\right)(t)\right\|\le \text{const.}\, t^{\gamma+(\bar{\alpha}+1)\bar{l}}(\ln t)^{\max(j_0,j_-)}\sigma_p(t)e^{q(t)}. \tag{3.93}$$

Only the construction of the "particular" solutions has to be changed in order to get these estimates.

The author conjectures that it is possible to change $\bar{r}$ to r defined in §2 so that all statements made should hold with r instead of $\bar{r}$.

4. Nonlinear problems. We consider problems of the form

$$y'=t^{\alpha}f(t,y), \qquad 1\le t<\infty, \quad \alpha\in N_0, \tag{4.1}$$

$$y\in C([1,\infty]), \tag{4.2}$$

$$b(y(1),y(\infty))=0. \tag{4.3}$$

We define for $a\in R^n$, $x,\bar{t}\in R$:

$$S_x(a)=\{y\in R^n \mid \|y-a\|\le x\}, \tag{4.4}$$

$$C_x(\bar{t},a)=\left\{(t,y)\in R^{n+1} \mid t\ge\bar{t}, y\in S_x(a)\right\} \tag{4.5}$$

and assume that

$$f,f_y\in C_{lip}(C_x(1,y(\infty))) \tag{4.6}$$

for some $x>0$ sufficiently large.

From (4.2) and (4.1) we conclude that

$$f(\infty,y(\infty))=0. \tag{4.7}$$

(4.7) is a (nonlinear) system of equations from which $y(\infty)$ can be calculated as a solution manifold $y(\infty)=y_\infty(\mu)$, $\mu\in S\subset R^{n_1}$, $n_1\leq n$, if the problem (4.1), (4.2), (4.3) admits a solution. The dimension of this manifold—n_1—is determined a priori if we require for a solution point y_∞ that

$$\operatorname{rank}\big(f_y(\infty,y_\infty)\big)=n-n_1 \tag{4.8}$$

$$\operatorname{rank}\big(f_y(\infty,y)\big)\leq n-n_1 \quad \text{for } y\in\mathring{S}_x(y_\infty),\quad x>0. \tag{4.9}$$

Then there is a $x_1>0$ and a n_1-dimensional manifold $y_\infty(\mu)$ which fulfills the equation

$$f(\infty,y_\infty(\mu))\equiv 0 \quad \text{for } \mu\in S\subset R^{n_1}, \tag{4.10}$$

$$y_\infty(\mu)\in\mathring{S}_{x_1}(y_\infty). \tag{4.11}$$

From (4.8) we conclude that n_1 equals the geometrical multiplicity of the eigenvalue zero of the matrix $f_y(\infty,y_\infty)$.

However, as practical examples point out, the assumption that $f_y(\infty,y)$ does not decrease its rank in y_∞ is too strong and therefore we regard n_1 as a priori unknown but obviously the solution of the equation (4.1) determines n_1 for a given problem.

Now we define

$$A(t,\mu)=f_y(t,y_\infty(\mu)) \tag{4.12}$$

and require that A fulfills (3.4), (3.5) so that

$$A(t,\mu)=\sum_{i=0}^{\infty} A_i(\mu)t^{-i} \quad \text{for } t\geq\delta. \tag{4.13}$$

We transform $A_0(\mu)$ to its Jordan canonical form $J_0(\mu)$

$$A_0(\mu)=E(\mu)J_0(\mu)E^{-1}(\mu) \tag{4.14}$$

and introduce z as new dependent variable

$$E(\mu)z=y-y_\infty(\mu) \tag{4.15}$$

getting

$$z'=t^\alpha J(t,\mu)z+t^\alpha g(z,t,\mu), \tag{4.16}$$

$$z(\infty)=0. \tag{4.17}$$

Here

$$J(t,\mu)=E^{-1}(\mu)A(t,\mu)E(\mu) \tag{4.18}$$

and

$$g(z,t,\mu)=E^{-1}(\mu)f(t,E(\mu)z+y_\infty(\mu))-J(t,\mu)z \tag{4.19}$$

hold. The perturbation g fulfills the estimates

$$\|g(z,t,\mu)\|\leq C_1(\mu)\big(\|f(t,y_\infty(\mu))\|+\|z\|^2\big), \tag{4.20}$$

$$\|g(z_1,t,\mu)-g(z_2,t,\mu)\|\leq C_2(\mu)\big(\|z_1\|+\|z_2\|\big)\|z_1-z_2\| \tag{4.21}$$

where $C_i(\mu)$ depend on the Lipschitz-constants of f,f_y on $C_x(1,y_\infty(\mu))$.

We restrict μ to subsets $\tilde{S}\subset S$ which are defined as follows.

1) The projections onto the direct sums of invariant subspaces of $J_0(\mu)$, which belong to eigenvalues with positive, zero and negative real part are constant for $\mu\in\tilde{S}$.

Moreover, the projections onto the invariant subspaces of $J_0(\mu)$ are constant for $\mu\in\tilde{S}$, therefore $r_+, r_0, r_-, \bar{r}$ are defined for $J_0(\mu)$ as in the last chapters and are independent of $\mu\in\tilde{S}$.

2) $y_\infty(\mu)$, $E(\mu)$, $E^{-1}(\mu)$ are continuous for $\mu\in\tilde{S}$.

Let $\phi(t,\mu)$ be the fundamental matrix of the (homogeneous) Problem (4.16).

3) The same columns $\varphi_i(t,\mu)$ of the matrix $\phi(t,\mu)$ fulfill

$$\|\varphi_i(t,\mu)\|\le C_i(\mu)t^{-(\alpha+1)\bar{r}-\varepsilon_1(\mu)}(\ln t)^{j_i},\qquad \varepsilon_1(\mu)>0 \tag{4.22}$$

for all $\mu\in\tilde{S}$. Therefore there is a projection matrix $\hat{D}_0$ independent of μ in $\tilde{S}$ so that

$$\|\phi(t,\mu)(\hat{D}_0+D_-)\|\le C(\mu)t^{-(\alpha+1)\bar{r}-\varepsilon_1(\mu)}(\ln t)^{j_0}\quad \text{for } \mu\in\tilde{S} \tag{4.23}$$

holds. Let $\hat{r}_0$ be the number of 1's in the main diagonal of $\hat{D}_0$.

We require f to fulfill

$$\|f(t,y_\infty(\mu))\|\le C(\mu)t^{-2(\alpha+1)\bar{r}-\varepsilon_2(\mu)},\qquad \varepsilon_2(\mu)>0\quad \text{for } \mu\in\tilde{S}. \tag{4.24}$$

For $\mu\in\tilde{S}$, $\xi\in C^n$ fixed we set

$$(\psi(z,\mu))(t)=\phi(t,\mu)(\hat{D}_0+D_-)\xi+(Hg(z,\cdot,\mu))(t), \tag{4.25}$$

where H is as in (3.70) (with $E=I$) and regard $\psi(\cdot,\mu)$ as an operator on the Banach space

$$\begin{aligned}&\left(A_{\varepsilon,\delta}=\left\{z\,|\,z(t)=Z(t)t^{-(\alpha+1)\bar{r}-\varepsilon}(\ln t)^{j_0},\ Z\in C_b([\delta,\infty))\right\},\ \|z\|_\varepsilon=\|Z\|_{[\delta,\infty]}\right),\\ &\qquad\qquad \delta\ge 1,\quad 0<\varepsilon=\min(\varepsilon_1,\varepsilon_2).\end{aligned} \tag{4.26}$$

Every fixed point z of $\psi(\cdot,\mu)$ establishes a solution for all $\xi\in C^n$. At first ψ maps $A_{\varepsilon,\delta}$ on $A_{\varepsilon,\delta}$ for δ sufficiently large because of (4.20), (4.23), (4.24), (3.69), (3.70).

Now we take a sufficiently large sphere $S_\varepsilon(\rho)$ in $A_{\varepsilon,\delta}$ with center $\phi(\cdot,\mu)(\hat{D}_0+D_-)\xi$ and radius ρ and prove the contraction property of ψ on $S_\varepsilon(\delta)$

$$\begin{aligned}\|\psi(z_1,\mu)-\psi(z_2,\mu)\|_\varepsilon&=\|H(g(z_1,\cdot,\mu)-g(z_2,\cdot,\mu))\|_\varepsilon\\ &\le \text{const.}(\mu)\cdot\rho\delta^{-\varepsilon}(\ln\delta)^{2j_0}\|z_1-z_2\|_\varepsilon\end{aligned} \tag{4.27}$$

for $z_1, z_2\in S_\varepsilon(\rho)$ because (4.21), (3.69) and (3.70) hold.

Moreover, if $z\in S_\varepsilon(\rho)$ then

$$\begin{aligned}\|\psi(z,\mu)-\phi(\cdot,\mu)(\hat{D}_0+D_-)\xi\|_\varepsilon&=\|Hg(z,\cdot,\mu)\|_\varepsilon\\ &\le(\text{const.}(\mu)+\rho)^2\cdot\delta^{-\varepsilon}(\ln\delta)^{2j_0}.\end{aligned} \tag{4.28}$$

Therefore $\psi(z,\mu)\in S_\varepsilon(\rho)$ if δ is sufficiently large and from (4.27), (4.28) we conclude that $\psi(\cdot,\mu)$ has a unique fixed point $z\in S_\varepsilon(\rho)\subset A_{\varepsilon,\delta}$ for δ sufficiently large. The construction of H implies that $(\hat{D}_0+D_-)P(\delta,\mu)^{-1}z(\delta)=\delta^{D(\mu)}e^{Q(\delta,\mu)}(\hat{D}_0+D_-)\xi$. Therefore, for fixed $\mu\in\tilde{S}$, we have constructed a $\hat{r}_0+r_-$ dimensional solution manifold in $A_{\varepsilon,\delta}$ for δ sufficiently large but fixed whenever $(\hat{D}_0+D_-)\xi$ varies in a compact set $K\subset C^{\hat{r}_0+r_-}$. In order to get more information on the asymptotic behavior of the solution we now treat the important case:

$$f(t,y_\infty(\mu))\equiv 0\quad \text{for } t\ge\delta,\quad \mu\in\tilde{S}, \tag{4.29}$$

and

$$\|\phi(t,\mu)(\hat{D}_0+D_-)\|=p(t,\mu)e^{q(t,\mu)}t^{\beta(\mu)}(\ln t)^{j_0},\qquad p\in C([1,\infty]) \tag{4.30}$$

where

$$q(t,\mu)\to -\infty \quad \text{for } t\to\infty \quad \text{and } \mu\in\tilde{S} \tag{4.31}$$

holds.

We define

$$\sigma(t,\mu)=e^{q(t,\mu)}t^{\beta(\mu)}(\ln t)^{j_0} \tag{4.32}$$

and set

$$(\bar{\psi}(z,\mu))(t)=\phi(t,\mu)(\hat{D}_0+D_-)\xi+(\bar{H}g(z,\cdot,\mu))(t), \qquad \xi\in C^n. \tag{4.33}$$

$\bar{H}$ is constructed according to §3 (with $E=I$) so that inhomogeneities f which decay as $\sigma^2(t,\mu)$ produce a particular solution $\bar{H}f$ which decays as $t^{(\alpha+1)\bar{l}}\sigma^2(t,\mu)(\ln t)^{\max(j_0,j_-)}$. We regard $\bar{\psi}$ as an operator on the Banach space

$$\left(A_{\sigma,\delta}=\{u \mid u=\sigma(t,\mu)U,\ U\in C_b([\delta,\infty))\},\ \|u\|_\sigma=\|U\|_{[\delta,\infty]}\right). \tag{4.34}$$

The construction mapping theorem, employed as in the case of algebraic decay, assures the existence of a (locally) unique fixed point z in $A_{\sigma,\delta}$.

From (4.33) we conclude

$$\|z(t)-\phi(t,\mu)(\hat{D}_0+D_-)\xi\|\le C(\mu)t^{(\alpha+1)\bar{l}}\sigma^2(t,\mu)(\ln t)^j, \tag{4.35}$$

where $j=\max(j_0,j_-)$.

It is easy to check that $\psi(\cdot,\mu)$ is also a contraction in a sphere around $\phi(\cdot,\mu)(\hat{D}_0+D_-)\xi$ in $A_{\sigma,\delta}$. The uniqueness of the fixed point assures that

$$\bar{H}g(z,\cdot,\mu)\in A_{\sigma,\delta} \tag{4.36}$$

for every fixed point z of $\psi(\cdot,\mu)$ in $A_{\varepsilon,\delta}$. Therefore

$$(\bar{\psi}(z,\mu))(t)=\phi(t,\mu)(\hat{D}_0+D_-)\xi+((H-\bar{H})g(z,\cdot,\mu))(t)+(\bar{H}g(z,\cdot,\mu))(t) \tag{4.37}$$

holds. Because Hg and $\bar{H}g$ are particular solutions for fixed $z\in A_{\sigma,\delta}$ we get

$$((H-\bar{H})g(z,\cdot,\mu)(t))=\phi(t,\mu)(\hat{D}_0+D_-)\gamma(z), \qquad \gamma(z)\in C^n. \tag{4.38}$$

Choosing

$$\zeta=\xi+\gamma(z) \tag{4.39}$$

assures that every fixed point of $\psi(z,\mu)$ is also a fixed point of $\bar{\psi}(z,\mu)$ (with ζ different to ξ) and vice versa.

In general our perturbation approach does not give us all $C([\delta,\infty])$ solutions of the problem (4.1), (4.2), (4.3); it only gives us all solutions, which decay at least as fast as $t^{-(\alpha+1)\bar{r}-\varepsilon}$ where $\varepsilon>0$. This is illustrated by the problem

$$y'=y^2, \qquad \delta\le t<\infty, \tag{4.40}$$

$$y\in C([\delta,\infty]) \tag{4.41}$$

which has the solution manifold $y=-1/(t+c)$, $c>-\delta$ and $y\equiv 0$. Our approach gives

$$y_\infty=0, \quad \frac{\partial f}{\partial y}(y_\infty)=0, \quad \bar{r}=1 \tag{4.42}$$

and therefore the fixed point equation

$$y=Hy^2 \tag{4.43}$$

results. In $A_{\varepsilon,\delta}$ the only solution of (4.43) is $y\equiv 0$, which is the only solution of (4.40) decaying faster than t^{-1}.

If $f_y(t,y_\infty(\mu))$ is not analytic at $t=\infty$ but if it fulfills a relation of the form (3.74) then $\psi_-^0(t,\mu)(D_-+\hat{D}_0)$ (resp. $\psi(g(z_1,\cdot,\mu))$) defined in §4 have to be used instead of $E(\mu)\phi(t,\mu)(\hat{D}_0+D_-)$ (resp. $E(\mu)Hg(z,\cdot,\mu)$). The results do not change.

The following theorem follows immediately.

THEOREM 4.1. *Let f, f_y fulfill a uniform Lipschitz condition in y on $[1,\infty]$. Then the problem* (4.1), (4.2), (4.3) *where* (4.13) *holds asymptotically has a solution $y\equiv y(\cdot,(\hat{D}_0+D_-)\xi,\mu)\equiv E(\mu)z(\cdot,(\hat{D}_0+D_-)\xi,\mu)+y_\infty(\mu)$ for every root $((\hat{D}_0+D_-)\xi,\mu)$ of the equation*

$$b\big(E(\mu)z\big(1,\big(\hat{D}_0+D_-\big)\xi,\mu\big)+y_\infty(\mu),y_\infty(\mu)\big)=0$$

where $(\xi,\mu)\in C^n\times\tilde{S}$ and $b: R^n\to R^{\hat{r}_0+r_-+n_1}$. Here $z(t,(\hat{D}_0+D_-)\xi,\mu)$ denote the continuated fixed points of $\psi(\cdot,\mu)$ with $\xi\in C^n$. On the other hand, if the boundary value problem (4.1), (4.2), (4.3) *has a solution y, so that $y-y(\infty)\in A_{\varepsilon,1}$ for some $\varepsilon>0$, then there is a $\mu\in R^{n_1}$ and a $(\hat{D}_0+D_-)\xi$, $\xi\in C^n$ so that $z:=E^{-1}(\mu)(y-y_\infty)$ has the asymptotic expansion* (4.35).

5. Case studies. In this section problems from fluid dynamics and thermodynamics described by boundary value problems on infinite intervals are analysed. These problems are represented by a nonautonomous system of nonlinear differential equations. An autonomous problem, namely von Karman's swirling flow problem, has been investigated by Markowich (1982a). The following equation represents a model for viscous flow past a solid at low Reynolds number:

$$u''+\frac{k}{x}u'+\alpha uu'+\beta(u')^2=0,\qquad \delta\le x<\infty,\quad \alpha>0, \tag{5.1}$$

$$u(\delta)=0,\quad \delta>0,\qquad u(\infty)=U>0. \tag{5.2}$$

The parameters and variables are described by Cohen, Fokas and Lagerström (1978). The transformation

$$y_1=u,\quad y_2=u',\quad y=(y_1,y_2)^T \tag{5.3}$$

takes (5.1), (5.2) into

$$y'=\begin{bmatrix} y_2 \\ -\frac{k}{x}y_2-\alpha y_1y_2-\beta y_2^2\end{bmatrix}=f(x,y), \tag{5.4}$$

$$\text{(a)}\quad [1,0]y(\delta)=0,\qquad \text{(b)}\quad [1,0]y(\infty)=U, \tag{5.5}$$

$$y\in C([\delta,\infty]). \tag{5.6}$$

We get $y_\infty=y_\infty(U)=\binom{U}{0}$ and calculate

$$f_y(x,y_\infty(U))=\underbrace{\begin{bmatrix}0 & 1\\ 0 & -\alpha U\end{bmatrix}}_{A_0(U)}+\frac{1}{x}\underbrace{\begin{bmatrix}0 & 0\\ 0 & -k\end{bmatrix}}_{A_1}. \tag{5.7}$$

$A_0(U)$ has the distinct eigenvalues 0 and $-\alpha U$ and we get

$$J_0(U)=\begin{bmatrix}0 & 0\\ 0 & -\alpha U\end{bmatrix},\qquad E(U)=\begin{bmatrix}1 & -\frac{1}{\alpha U}\\ 0 & 1\end{bmatrix}. \tag{5.8}$$

The transformation (4.15) with $\mu = U$ results in the system

$$z' = \left(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & -\alpha U \end{bmatrix}}_{J_0(U)} + \frac{1}{x} \underbrace{\begin{bmatrix} 0 & -\frac{k}{\alpha U} \\ 0 & -k \end{bmatrix}}_{J_1(U)} \right) z + g(z, U), \tag{5.9}$$

$$z(\infty) = 0. \tag{5.10}$$

The homogeneous problem $z_h' = (J_0(U) + \frac{1}{x} J_1(U)) z_h$ has a fundamental matrix $\phi(x, U)$ of the form

$$\phi(x, U) = P(x, U) \begin{bmatrix} 1 & 0 \\ 0 & e^{-\alpha U x} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & x^{-k} \end{bmatrix}, \tag{5.11}$$

where

$$P(x, U) = I + O(x^{-1}) \tag{5.12}$$

holds. This follows from Theorem 3.1 and from the algorithm for the calculation of the coefficients. Moreover,

$$\hat{D}_0 = 0, \qquad D_- = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{5.13}$$

holds and the fixed point equation

$$z(t) = P(t, U) \begin{bmatrix} 0 \\ e^{-\alpha U x} x^{-k} \xi \end{bmatrix} + (\bar{H} g(z, U))(t), \qquad t \in R \tag{5.14}$$

results.

$g(z, U)$ fulfills

$$\| g(z, U) \| \leq \text{const.}(U) \cdot \| z \|^2 \tag{5.15}$$

because $f(x, y_\infty(U)) = 0$ holds.

Section 4 assures the existence of solutions $z(\cdot, \xi, U) \in A_{\varepsilon, \bar{\delta}} = \{u \mid u(x) = x^{-1-\varepsilon} \ln x U(x),\ U \in C_b([\bar{\delta}, \infty))\}$ and from (4.35) we conclude:

$$\left\| z(x, \xi, U) - P(x, U) \begin{bmatrix} 0 \\ e^{-\alpha U x} x^{-k} \xi \end{bmatrix} \right\| \leq C(U) x^{1-2k} e^{-2\alpha U x} (\ln x)^2. \tag{5.16}$$

This estimate can be improved, so that its right-hand side is of the order $x^{-2k} e^{-2\alpha U x}$. Resubstituting we get

$$u(x, \xi, U) = U - \frac{1}{\alpha U} (1 + O(x^{-1})) e^{-\alpha U x} x^{-k} \xi + O(e^{-2\alpha U x} x^{-2k}), \qquad x \to \infty. \tag{5.17}$$

Since $U \in R^+$ is given (5.5a) has to be used in order to determine $\xi \in R$ and the problem is well posed concerning the number of conditions as $x = \delta$ and $x = \infty$.

The second problem is a similarity equation for a combined forced and free convection flow over a horizontal plate (see Schneider (1979))

$$\begin{aligned} &\text{(a)} \quad 2f''' + ff'' + kxg = 0, \\ &\text{(b)} \quad 2g' + fg = 0, \end{aligned} \qquad 0 \leq x < \infty, \tag{5.18}$$

$$\text{(a)} \quad f(0) = f'(0) = 0, \quad g(0) = 1, \qquad \text{(b)} \quad f'(\infty) = 1, \quad g(\infty) = 0. \tag{5.19}$$

The variables and the parameter k are explained in Schneider (1979). For simplicity we have set the Prandtl number to 1. Because of (5.19b) we substitute

$$f(x)=x+h(x) \tag{5.20}$$

and get the new problem

$$\begin{aligned} &\text{(a)} && 2h'''+(x+h)h''+kxg=0, \\ &\text{(b)} && 2g'+(x+h)g=0, \end{aligned} \qquad 0\le x<\infty, \tag{5.21}$$

$$\text{(a)}\quad h(0)=0,\quad h'(0)=-1,\quad g(0)=1, \qquad \text{(b)}\quad h'(\infty)=0,\quad g(\infty)=0. \tag{5.22}$$

Substituting

$$y_1=h,\quad y_2=h',\quad y_3=h'',\quad y_4=g,\quad y=(y_1,y_2,y_3,y_4)^T \tag{5.23}$$

we get the system

$$y'=x\begin{bmatrix} \dfrac{y_2}{x} \\ \dfrac{y_3}{x} \\ -\dfrac{1}{2}\left(1+\dfrac{y_1}{x}\right)y_3-\dfrac{k}{2}y_4 \\ -\dfrac{1}{2}\left(1+\dfrac{y_1}{x}\right)y_4 \end{bmatrix}=xf(x,y), \qquad 0\le x<\infty, \tag{5.24}$$

$$\text{(a)}\quad \begin{bmatrix} 1&0&0&0\\0&1&0&0\\0&0&0&1 \end{bmatrix}y(0)=\begin{bmatrix} 1\\-1\\1 \end{bmatrix}, \qquad \text{(b)}\quad \begin{bmatrix} 0&1&0&0\\0&0&0&1 \end{bmatrix}y(\infty)=\begin{bmatrix} 0\\0 \end{bmatrix}. \tag{5.25}$$

We only admit solutions fulfilling $h(\infty)\in R$, therefore we require that

$$y\in C([0,\infty]) \tag{5.26}$$

holds. From (5.23), (5.24) we conclude

$$y_\infty=y_\infty(h_\infty)=(h_\infty,0,0,0), \qquad h_\infty=h(\infty)\in R. \tag{5.27}$$

We calculate:

$$f_y(x,y_\infty(h_\infty))=\underbrace{\begin{bmatrix} 0&0&0&0\\0&0&0&0\\0&0&-\frac{1}{2}&-\frac{k}{2}\\0&0&0&-\frac{1}{2} \end{bmatrix}}_{A_0}+\frac{1}{x}\underbrace{\begin{bmatrix} 0&1&0&0\\0&0&1&0\\0&0&-\frac{h_\infty}{2}&0\\0&0&0&-\frac{h_\infty}{2} \end{bmatrix}}_{A_1(h_\infty)} \tag{5.28}$$

and

$$E\equiv E(h_\infty)=\operatorname{diag}\left(1,1,1,-\frac{2}{k}\right). \tag{5.29}$$

The substitution $E(h_\infty)z=y-y_\infty(h_\infty)$ gives the system

$$z'=x\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&-\frac{1}{2}&1\\0&0&0&-\frac{1}{2}\end{bmatrix}+\frac{1}{x}\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&-\frac{h_\infty}{2}&0\\0&0&0&-\frac{h_\infty}{2}\end{bmatrix}z+xg(z,h_\infty). \tag{5.30}$$

J_0 has the eigenvalues $\lambda_1=0$ with algebraic and geometric multiplicity 2 and $\lambda_2=-1/2$ with algebraic multiplicity 2 and geometric multiplicity 1. Because $f(x,y_\infty(h_\infty))=0$ holds we get

$$\|g(z,h_\infty)\|\le C(h_\infty)\|z\|^2. \tag{5.31a}$$

We have to set up the fundamental matrix $\phi(x,h_\infty)$ of the system

$$z_h'=x\left(J_0+\frac{1}{x}J_1(h_\infty)\right)z_h,\qquad z_h=\left(z_h^{(1)},z_h^{(2)},z_h^{(3)},z_h^{(4)}\right). \tag{5.31b}$$

Because of the simple structure of J_0+J_1/x we do not have to apply the algorithm of §3; we can proceed in the following way.

The last equation of (5.31) is

$$z_h^{(4)'}=x\left(-\frac{1}{2}-\frac{h_\infty}{2}\frac{1}{x}\right)z_h^{(4)}. \tag{5.32}$$

It can be integrated at once giving

$$z_h^{(4)}=e^{-x^2/4-h_\infty x/2}c,\qquad c\in R. \tag{5.33}$$

Setting $c=1$ we have to find a particular solution of

$$z_h^{(3)'}=x\left(-\frac{1}{2}-\frac{h_\infty}{2}\frac{1}{x}\right)z_h^{(3)}+xz_h^{(4)}. \tag{5.34}$$

We take

$$z_h^{(3)}=e^{-x^2/4-h_\infty x/2}\frac{x^2}{2}\left(1-\frac{1}{x^2}\right). \tag{5.35}$$

Integrating

$$z_h^{(2)'}=z_h^{(3)} \tag{5.36}$$

we find

$$z_h^{(2)}=e^{-x^2/4-h_\infty x/2}xO(1). \tag{5.37}$$

Analogously, we integrate

$$z_h^{(1)'}=z_h^{(2)} \tag{5.38}$$

and get

$$z_h^{(1)}=e^{-x^2/4-h_\infty x/2}O(1). \tag{5.39}$$

The fourth column of $\phi(x,h_\infty)$ can be chosen as $(z_h^{(1)},z_h^{(2)},z_h^{(3)},z_h^{(4)})^T$. In order to get the third column we set $c=0$ in (5.33) and proceed as we did.

Finally, we get the following fundamental matrix

$$\phi(x,h_\infty)=\begin{bmatrix} 1 & x & e^{-x^2/4-h_\infty x/2}O(x^{-2}) & e^{-x^2/4-h_\infty x/2}O(1) \\ 0 & 1 & e^{-x^2/4-h_\infty x/2}O(x^{-1}) & e^{-x^2/4-h_\infty x/2}xO(1) \\ 0 & 0 & e^{-x^2/4-h_\infty x/2} & e^{-x^2/4-h_\infty x/2}x^2\left(1-\frac{1}{x^2}\right) \\ 0 & 0 & 0 & e^{-x^2/4-h_\infty x/2} \end{bmatrix} \tag{5.40}$$

which we can write as in Theorem 3.1

$$\phi(x,h_\infty)=P(x,h_\infty)x^D e^{Q(x,h_\infty)}, \tag{5.41}$$

where

$$P(x,h_\infty)=\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}+\frac{1}{x}O(1), \tag{5.42}$$

$$D=\operatorname{diag}(1,x,1,x^2), \tag{5.43}$$

$$Q(x,h_\infty)=\operatorname{diag}\left(1,1,e^{-x^2/4-h_\infty x/2}x,e^{-x^2/4-h_\infty x/2}x\right) \tag{5.44}$$

hold.

Similarily to the first example, we conclude that there are solutions $z(x,\xi_1,\xi_2,h_\infty)$ in $A_{\varepsilon,\delta}$, which is now the space of all functions in $C([\delta,\infty])$ which decay at least as $x^{-4-\varepsilon}\ln x$, $\varepsilon>0$ because $\bar{r}=2$ and $\alpha=1$ holds. Moreover,

$$z(x,\xi_1,\xi_2 h_\infty)=P(x,h_\infty)\begin{bmatrix} 0 \\ 0 \\ x e^{-x^2/4-h_\infty x/2}x\xi_1 \\ x^2 e^{-x^2/4-h_\infty x/2}x\xi_2 \end{bmatrix}+O\left(x^4 e^{-x^2/2-h_\infty x}\right) \tag{5.45}$$

with h_∞, ξ_1, $\xi_2\in R$ follows because the exponential factor $e^{-x^2/2-h_\infty x}$ does not appear in the fundamental matrix.

So we get the asymptotic expansions

$$f(x)=x+h_\infty+O(1)e^{-x^2/4-h_\infty x/2}x+O\left(x^4 e^{-x^2/2-h_\infty x}\right), \tag{5.46}$$

$$g(x)=-\frac{2}{k}e^{-x^2/4-h_\infty x/2}\xi_2+O\left(x^4 e^{-x^2/2-h_\infty x}\right) \tag{5.47}$$

where the $O(1)$ in (5.46) depends linearly on ξ_1 and ξ_2.

The constants $(h_\infty,\xi_1,\xi_2)\in R^3$ have to be determined from the three initial conditions (5.25a).

The third problem to be analysed is the well-known Falkner-Skan equation (see, for example, Schlichting (1951))

$$f'''+f''+(1-f'^2)=0, \tag{5.48}$$

$$f'(\infty)=1. \tag{5.49}$$

We do not pose any initial condition because we look for a solution manifold.

Because of (5.49) we substitute

$$f(x)=x+g(x),\quad y_1=g,\quad y_2=g',\quad y_3=g'', \tag{5.50}$$

$$y=(y_1,y_2,y_3)^T \tag{5.51}$$

and get the system

$$y'=x\begin{bmatrix} \dfrac{y_2}{x} \\ \dfrac{y_3}{x} \\ -y_3-\dfrac{y_1y_3}{x}+\dfrac{2y_2}{x}+\dfrac{y_2^2}{x} \end{bmatrix}=xf(x,y),\qquad x\geq\delta. \tag{5.52}$$

Moreover, we require

$$y\in C([\delta,\infty]) \tag{5.53}$$

so that

$$y(\infty)=(g_\infty,0,0)^T,\qquad g_\infty:=g(\infty) \tag{5.54}$$

holds. We calculate

$$f_y(x,y_\infty(g_\infty))=\underbrace{\begin{bmatrix}0&0&0\\0&0&0\\0&0&-1\end{bmatrix}}_{J_0}+\frac{1}{x}\underbrace{\begin{bmatrix}0&1&0\\0&0&1\\0&2&-g_\infty\end{bmatrix}}_{J_1(g_\infty)}. \tag{5.55}$$

The linearized system is therefore

$$z_h'=x\left(J_0+\frac{1}{x}J_1(g_\infty)\right)z_h. \tag{5.56}$$

Because J_0 has the eigenvalue 0 with algebraic multiplicity 2 we apply the theory developed in §3.

At first we split up the system (5.56) by the transformation

$$z_h=P(x,g_\infty)u,\quad u=(u_1,u_2,u_3)^T,\quad P(x,g_\infty)\sim I+\sum_{i=1}^{\infty}P_i(g_\infty)x^{-i} \tag{5.57}$$

to get subsystems whose leading matrices have the only eigenvalue 0 and -1, respectively. (5.57) gives a system of the form

$$u'=xB(x,g_\infty)u,\qquad B(x,g_\infty)\sim J_0+\sum_{i=1}^{\infty}B_i(g_\infty)x^{-i}. \tag{5.58}$$

From Wasow (1965) we conclude that

$$P_i=\begin{bmatrix}0&0&p_{i1}\\0&0&p_{i2}\\p_{i3}&p_{i3}&0\end{bmatrix},\qquad B_i=\begin{bmatrix}b_{i1}&b_{i2}&0\\b_{i3}&b_{i4}&0\\0&0&b_{i5}\end{bmatrix} \tag{5.59}$$

holds and that the recursion

$$J_0P_i-P_iJ_0=\sum_{s=0}^{i-1}(P_sB_{i-s}-J_{i-s}P_s)-(i-2)P_{i-2},\qquad i>0 \tag{5.60}$$

with the last term absent for $i<2$ and $J_k=0$ for $k>1$ holds.

From the investigation of the perturbed system we know that only the coefficients B_0, B_1, B_2, B_3 and B_4 influence the asymptotic behaviour of the fundamental matrix because $(\alpha+1)\bar{r}=4$ for our example. (5.59) and (5.60) give

$$
\begin{aligned}
&B_1=\begin{bmatrix}0&1&0\\0&0&0\\0&0&-g_\infty\end{bmatrix}, \qquad B_2=\begin{bmatrix}0&0&0\\0&2&0\\0&0&-2\end{bmatrix},\\
&B_3=\begin{bmatrix}0&0&0\\0&-2g_\infty&0\\0&0&-2g_\infty\end{bmatrix}, \qquad B_4=\begin{bmatrix}0&0&0\\0&-2+2g_\infty^2&0\\0&0&b_{45}\end{bmatrix}.
\end{aligned}
\tag{5.61}
$$

We do not have to know b_{45} explicitly because it does not influence the behaviour of the solution of

$$
u_3'=x\left(-1-\frac{g_\infty}{x}-\frac{2}{x^2}+\frac{2g_\infty}{x^3}+\cdots\right)u_3. \tag{5.62}
$$

From §3 we get

$$
u_3=p_3(x,g_\infty)e^{-x^2/2-g_\infty x}x^{-2}, \qquad p_3(x,g_\infty)\sim 1+p_{31}(g_\infty)x^{-1}+p_{32}(g_\infty)x^{-2}+\cdots. \tag{5.63}
$$

Moreover, we get

$$
\begin{pmatrix}u_1\\u_2\end{pmatrix}'=\left(\begin{bmatrix}0&1\\0&0\end{bmatrix}+\frac{1}{x}\begin{bmatrix}0&0\\0&2\end{bmatrix}+\frac{1}{x^2}\begin{bmatrix}0&0\\0&-2g_\infty\end{bmatrix}+\frac{1}{x^3}\begin{bmatrix}0&0\\0&-2+2g_\infty^2\end{bmatrix}+\cdots\right)\begin{pmatrix}u_1\\u_2\end{pmatrix} \tag{5.64}
$$

because the leading term which comes from B_0 vanishes. Therefore the coefficient of x^{-3} does not influence the behaviour of u_1, u_2. It is sufficient to solve

$$
\begin{bmatrix}\tilde{u}_1\\\tilde{u}_2\end{bmatrix}'=\left(\begin{bmatrix}0&1\\0&0\end{bmatrix}+\frac{1}{x}\begin{bmatrix}0&0\\0&2\end{bmatrix}+\frac{1}{x^2}\begin{bmatrix}0&0\\0&-2g_\infty\end{bmatrix}\right)\begin{bmatrix}\tilde{u}_1\\\tilde{u}_2\end{bmatrix}. \tag{5.65}
$$

We get directly by integration

$$
\tilde{u}_2=x^2\exp\left(\frac{2g_\infty}{x}\right) \quad \text{and} \quad \tilde{u}_1=x^3O(1). \tag{5.66}
$$

Finally, we conclude

$$
f=x+g_\infty+p(x,g_\infty)e^{-x^2/2-g_\infty x}\xi+O\left(e^{-x^2-2g_\infty x}\right), \qquad p(x,g_\infty)\sim 1+\sum_{i=1}^{\infty}p_i(g_\infty)^{x^{-i}} \tag{5.67}
$$

with $\xi\in R$ (because we look for real solutions) and because there is no column of the fundamental matrix of (5.56) which contains the factor $e^{-x^2-2g_\infty x}$.

REFERENCES

[1] E. Coddington and N. Levinson (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.

[2] D. S. Cohen, A. Fokas and P. A. Lagerström (1978), *Proof of some asymptotic results for a model equation for low Reynolds number flow*, SIAM J. Appl. Math., 35, pp. 187–207.

[3] F. R. de Hoog and Richard 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.

[4] _____(1980b), *An approximation method for boundary value problems on infinite intervals*, Computing, 24, pp. 227–239.

[5] A. D. JEPSON (1980), *Asymptotic boundary conditions for ordinary differential equations*, Ph.D. Thesis, California Institute of Technology, Pasadena.

[6] 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.

[7] _____(1980b), *The von Karman swirling flows*, SIAM J. Appl. Math., 38, pp. 52–64.

[8] P. A. MARKOWICH (1980), *Randwertprobleme auf unendlichen Intervallen*, Dissertation, Technische Universitat Wien, Vienna.

[9] _____ (1982a), *Asymptotic analysis of von Karman flows*, SIAM J. Appl. Math., 42, pp. 549–557.

[10] _____(1982b), *A theory for the approximation of solution of boundary value problems on infinite intervals*, this Journal, 13, pp. 484–513.

[11] J. B. MCLEOD (1969), *Von Karman's swirling flow problem*, Arch. Rat. Mech. Anal., 1, pp. 99–102.

[12] H. SCHLICHTING (1951), *Grenzschicht Theorie*, Karlsruhe, Braun.

[13] W. SCHNEIDER (1979), *A similarity solution for combined forced and free convection flow over a horizontal plate*, Internat. J. Heat Mass Transfer, 22, pp. 1401–1406.

[14] W. WASOW (1965), *Asymptotic Expansion for Ordinary Differential Equations*, Series in Pure and Applied Mathematics XIV, John Wiley, New York.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0003 $01.25/0

COCURRENT FLOW IN A BUBBLE-COLUMN REACTOR AT HIGH PÉCLET NUMBERS*

J. J. SHEPHERD†

Abstract. We consider a nonlinear system of equations that models the action of a bubble-column reactor, and which becomes a singular perturbation problem in the limit of infinitely large Péclet numbers. For such parameter regimes, we establish existence and uniqueness results for this problem, and indicate methods by which successively closer approximations to its solution may be obtained.

1. Introduction. Over a number of years, various models have been developed that attempt to describe the action of a bubble-column reactor, in which the chemical reaction desired is induced by allowing bubbles of a reactive gas (or mixture of gases) to rise vertically through a column of the appropriate liquid. The mathematical analysis of such a system is complicated, reflecting the complex behaviour observed experimentally.

Earlier models of such a reactor treated the reaction between the gas and liquid phases as analogous to that occurring in a reactor involving contact between like phases. While the equations developed on this basis to model this situation were linear and readily solved, they ignored several effects that prove to be of significance. Most important of these is that removal of material from a gas bubble alters the size of this bubble appreciably, and hence its rate of rise in the liquid column. If we couple this with the effects of the pressure gradient in a tall column, we see that the residence time of bubbles in the region of greatest concentration of unreacted liquid phase reactant (i.e., near the entrance to the column) may be appreciably increased, with a marked effect on the concentration profiles of quantities along the column.

Thus, as pointed out by Deckwer [3], these effects must be included in any equations attempting to model such a reactor. In particular, the rise velocity of bubbles may not be regarded as constant, but becomes, in fact, one of the unknown quantities of the system to be determined. In the light of these considerations, Deckwer has proposed a system of equations describing steady-state concentrations in gas and liquid phases, as well as the bubble velocity, along the axis of the column. For the simplest case of a single gas reacting with and being absorbed by a simple liquid, and with the flow of gas bubbles and liquid being in the same axial direction (termed *cocurrent* flow), his equations modelling the situation may be written in dimensionless form as follows:

$$u'(z)=\alpha\beta(z)^{-1}u+\sigma_G x_0 F(x,y), \tag{1.1}$$

$$\varepsilon_G x''(z)-\left(2\alpha\varepsilon_G\beta(z)^{-1}+u\right)x'=-\sigma_G(1-x_0x)F(x,y), \tag{1.2}$$

$$\varepsilon_L y''-y'=Dy+\sigma_L\beta(z)(1+\alpha)^{-1}F(x,y) \tag{1.3}$$

on $0<z<1$, where

$$\beta(z)=1+\alpha(1-z), \tag{1.4}$$

and

$$F(x,y)=(1+\alpha)\mu^{-1}\beta(z)^{-1}y-x. \tag{1.5}$$

* Received by the editors February 9, 1981, and in revised form December 8, 1981.

† Department of Mathematics, Royal Melbourne Institute of Technology, P.O. Box 2476V, Melbourne 3001, Australia

Boundary conditions at the ends $z=0, 1$ of the column are given as

$$u(0)=1, \tag{1.6}$$

$$\varepsilon_G x'(0)-x(0)=-1, \tag{1.7}$$

$$\varepsilon_L y'(0)-y(0)=0, \tag{1.8}$$

$$x'(1)=0 \tag{1.9}$$

and

$$y'(1)=0. \tag{1.10}$$

In the above, as already noted, z is a length variable measured along the axis of the column and scaled with respect to the length of that column. $x(z)$ is the reduced mole fraction of the gas component of reactant in the gas phase (i.e., the mole fraction normalized with respect to its value x_0 at $z=0$, the entrance of the column), while $y(z)$ and $u(z)$ represent the dimensionless concentration of gas in the liquid phase and the superficial gas velocity along the column, respectively. ε_G and ε_L are the reciprocals of the gas and liquid phase Péclet numbers, while σ_G and σ_L are the gas and liquid phase Stanton numbers respectively. μ is a constant of proportionality for the rate of absorption of the gas in liquid, while D is the Damköhler number, which measures the rate of gas uptake in reaction relative to the total liquid convection rate. The constant α determines the (linear) pressure field through the relationship (1.4)—in most cases, large values of α signify tall bubble columns.

We will term the full boundary-value problem (1.1)–(1.10) Problem A. Obviously, the occurrence of the eight parameters ε_G, ε_L, σ_G, σ_L, α, μ, x_0 and D in this problem leads to an almost infinite number of possibilities when one considers the question of solutions of this system. Moreover, the literature available (see [3], [6] for example) on this system gives little scope for deciding on a regime of parameter values that is most commonly encountered in practice. While we appreciate that the parameter ranges (such as they are), indicated there, are relevant only to the numerical and experimental investigations of the writers concerned, they may be used as an indication for values to be adopted in a rigorous analysis of Problem A. On the basis of these results, we find that, typically (though not always), $\varepsilon_G \ll 1$. Such values might arise when the bubble entrance velocity is very high. Moreover, difficulties reported by Deckwer [3] in the numerical treatment of Problem A at high Péclet numbers leads us also to consider the case $\varepsilon_L \ll 1$. Apart from these assumptions, we regard all other parameters as positive order one quantities, while the physics of the situation lead to the restrictions $0<x_0<1$, $\mu \geq 1$ and $D \geq 0$ ($D=0$ corresponding to pure physical absorption).

Motivated by the above discussion, we propose, in this paper, to analyze Problem A by asymptotic methods in the limit as $\varepsilon_G, \varepsilon_L \to 0$. We make no assumptions about the relative rates of convergence to zero of ε_G and ε_L, but maintain their independence (on the basis of typical parameter values used in Shioya, Dang and Dunn [5], $\varepsilon_L=\sqrt{\varepsilon_G}$ might be feasible). Clearly, under this limiting process, Problem A is a singular perturbation problem, since the *reduced problem* obtained by formally setting $\varepsilon_G=\varepsilon_L=0$ is a first order initial-value problem and cannot satisfy two-point boundary conditions. Thus we anticipate the appearance of boundary layers; and since, on physical grounds, $u(z)>0$, the form of Problem A indicates that such layers will occur in a neighbourhood of $z=1$. On the basis of this, we expect the solution $U(z)$, $X(z)$ $Y(z)$ of the problem

$$U'(z)=\alpha\beta(z)^{-1}U+\sigma_G x_0 F(X,Y), \tag{1.11}$$

$$U(z)X'(z)=\sigma_G(1-x_0X)F(X,Y), \tag{1.12}$$

$$Y'(z)=-DY-\sigma_L\beta(z)(1+\alpha)^{-1}F(X,Y), \tag{1.13}$$

$$U(0)=1,\quad X(0)=1,\quad Y(0)=0, \tag{1.14}$$

to suitably approximate the solution of Problem A as $\varepsilon_G,\varepsilon_L\to 0$ for all z in $[0,1]$ except a neighbourhood of $z=1$. What is more, the Neumann-type boundary conditions (1.9), (1.10) at $z=1$ lead us to expect that, in fact, U,X,Y above approximate the solution to Problem A throughout *all* of $[0,1]$. This will in fact turn out to be the case (i.e., the boundary layers at $z=1$ are of higher order).

The problem (1.11)–(1.14) may be simplified somewhat. We may eliminate $F(X,Y)$ between (1.11) and (1.12), and by integrating and applying the boundary conditions (1.14), obtain the basic relationship

$$U(z)=\frac{(1-x_0)(1+\alpha)}{(1-x_0X)\beta(z)}, \tag{1.15}$$

while the differential equation (1.12) becomes

$$X'(z)=\sigma_G(1-x_0X)^2(1-x_0)^{-1}(1+\alpha)^{-1}\beta(z)F(X,Y). \tag{1.16}$$

We will term the problem (1.15), (1.16), (1.13) together with the last two of the boundary conditions (1.14) Problem B. In the sections that follow, we will consider the questions of existence of solutions to Problems A and B, and the relationship between these solutions when $\varepsilon_G,\varepsilon_L\to 0$. Our basic tool will the contraction mapping theorem [5, p. 26], which has the advantage of giving an iteration scheme with rapid convergence, at the price of sometimes indelicate estimates on the solutions involved.

2. Existence for Problem A. We here examine the existence-uniqueness question for Problem A by relating it to that for Problem B. Systems resembling Problem A have been dealt with by Chang [1] and later by Kelley [4] in considerable generality. While Chang's techniques might be adaptable to the case at hand, we will find it advantageous to adopt a more direct method of analysis, that exploits the accessibility of the structure of the equations of Problem A. Moreover, it gives us tighter estimates on the solutions considered, as well as keeping separate the distinct rates of convergence to zero of ε_G and ε_L.

We begin by supposing that Problem B has a solution $(U(z),X(z),Y(z))$, where U,X,Y are suitably continuously differentiable on $[0,1]$, and where $U(z)$ is positive there. Since Problem B is independent of ε_G and ε_L, these properties will hold as $\varepsilon_G,\varepsilon_L\to 0$. If we let (u,x,y) be the (proposed) solution of Problem A, we may change the dependent variables in Problem A to $(\zeta(z),\eta(z),\xi(z))$, defined by

$$(u,x,y)=(U+\xi,X+\eta,Y+\zeta), \tag{2.1}$$

and on the basis of the arguments of the previous section, seek solutions (ξ,η,ζ) that are small in some sense as $\varepsilon_G,\varepsilon_L\to 0$. By substituting and linearizing, we may now write Problem A in terms of (ξ,η,ζ) as

$$\xi'-\alpha\beta(z)^{-1}\xi=p\eta+q\zeta, \tag{2.2}$$

$$\varepsilon_G\eta''-a_1\eta'+b_1\eta+c_1\zeta+d_1\xi=O(\varepsilon_G)+Q(\eta,\zeta), \tag{2.3}$$

$$\varepsilon_L\zeta''-\zeta'+b_2\zeta+c_2\eta=O(\varepsilon_L). \tag{2.4}$$

with boundary conditions

$$\xi(0)=0, \tag{2.5}$$

$$\varepsilon_G\eta'(0)-\eta(0)=O(\varepsilon_G), \tag{2.6}$$

$$\varepsilon_L\zeta'(0)-\zeta(0)=O(\varepsilon_L), \tag{2.7}$$

$$\eta'(1)=O(1), \tag{2.8}$$

$$\zeta'(1)=O(1), \tag{2.9}$$

where the order symbols are uniform in z as $\varepsilon_G, \varepsilon_L \to 0$.

In the above, we have, explicitly,

$$a_1(z)=U(z)+\xi+O(\varepsilon_G), \tag{2.10}$$

while $p(z), q(z), b_1(z), b_2(z), c_1(z), c_2(z)$ and $d_1(z)$ are all functions that are continuous on [0, 1] and $O(1)$ there as $\varepsilon_G, \varepsilon_L \to 0$. $Q(\eta, \zeta)$ is a term that is quadratic in η and ζ, with coefficients that are $O(1)$ as $\varepsilon_G, \varepsilon_L \to 0$.

Consider now the above system for (ξ, η, ζ). We may construct an associated linear system by replacing $a_1(z)$ by $\bar{a}_1(z)=U(z)+\bar{\xi}+O(\varepsilon_G)$ and (ξ, η, ζ) by $(\bar{\xi}, \bar{\eta}, \bar{\zeta})$ in the right-hand sides of (2.2)–(2.4). By regarding this as a map from functions $(\bar{\xi}, \bar{\eta}, \bar{\zeta})$ into functions (ξ, η, ζ), we may view the solution of the original system (when it exists) as a fixed point of this map. A standard framework within which to apply fixed point theorems would be provided by converting this (linear) system to a system of integral equations. Unfortunately, while this may prove possible for given $\varepsilon_G, \varepsilon_L > 0$, the behaviour of such systems as $\varepsilon_L, \varepsilon_G \to 0$ is not well understood—at least to the degree desired here. The work of Chang [1] and others assumes that $\varepsilon_G = \varepsilon_L$, while boundary conditions of the type arising here are not considered.

To exploit some of these ideas, but to avoid the difficulties, we adopt the procedure below. We introduce functions $A(z), B(z), C(z), D(z)$, continuously differentiable on [0, 1], such that

$$\eta = A + Be^{-\chi_1}, \qquad \eta' = \varepsilon_G^{-1} a_1 Be^{-\chi_1} \tag{2.11}$$

and

$$\zeta = C + De^{-\chi_2}, \qquad \zeta' = \varepsilon_L^{-1} De^{-\chi_2}, \tag{2.12}$$

where χ_1 and χ_2 are defined by

$$\chi_1(z) = \varepsilon_G^{-1} \int_z^1 a_1(s)\, ds, \qquad \chi_2(z) = \varepsilon_L^{-1}(1-z) \tag{2.13}$$

respectively. Clearly, A, B, C and D also depend on ε_G and ε_L.

The forms assumed for η and ζ are motivated by the observation that, for ξ (or $\bar{\xi}$) small, $a_1(z)$ (or $\bar{a}_1(z)$) > 0 on [0, 1], so that the second-order differential operators for η and ζ on the left-hand sides of (2.3) and (2.4) are of a form standard to the singular perturbation literature. Boundary-value problems involving such operators have been extensively studied and the forms of solutions to such problems are well known, consisting of an "outer" solution and a "boundary layer correction" that is appreciable only near $z = 1$ (in this case). The forms assumed for η' and ζ' are a little different, and their adoption implies the compatibility conditions

$$A' + B'e^{-\chi_1} = 0, \tag{2.14}$$

$$C' + D'e^{-\chi_2} = 0. \tag{2.15}$$

By substituting for η and ζ from (2.11), (2.12) into the equations (2.2), (2.3), (2.4), and applying the conditions (2.14), (2.15), we obtain a first-order system for ξ, A, B, C and D as below:

$$\xi' - \alpha\beta(z)^{-1}\xi - pA - qC = pBe^{-\chi_1} + qDe^{-\chi_2}, \tag{2.16}$$

$$A' - b_1 a_1^{-1} A - a_1^{-1} d_1 \xi - a_1^{-1} c_1 C = a_1^{-1}\left[(b_1 + a_1')Be^{-\chi_1} + c_1 De^{-\chi_2} - r_1\right], \tag{2.17}$$

$$C' - b_2 C - c_2 A = c_2 Be^{-\chi_1} + b_2 De^{-\chi_2} - r_2, \tag{2.18}$$

$$B' + a_1^{-1}(b_1 + a_1')B = a_1^{-1} e^{\chi_1}\left[-b_1 A - c_1 C - d_1 \xi - c_1 De^{-\chi_2} + r_1\right], \tag{2.19}$$

$$D' + b_2 D = e^{\chi_2}\left[-c_2 A - c_2 Be^{-\chi_1} - b_2 C + r_2\right], \tag{2.20}$$

where $r_1(z)$, $r_2(z)$ are the right-hand sides from (2.3) and (2.4) respectively. Note that for $\xi = o(1)$ as $\varepsilon_G, \varepsilon_L \to 0$, $a_1(z) > 0$ on $[0, 1]$ when $U(z) > 0$ there.

Consider now the equations (2.16)–(2.18). From (2.6) we obtain

$$A(0) = O(\varepsilon_G) + O(1)B(0)e^{-\chi_1(0)}, \tag{2.21}$$

while, from (2.7),

$$C(0) = O(\varepsilon_L) + O(1)D(0)e^{-\chi_2(0)}. \tag{2.22}$$

We may now introduce a fundamental matrix, Φ, corresponding to the linear differential operator on the left-hand sides of (2.16)–(2.18) and we obtain, on integrating and noting (2.21), (2.22),

$$\begin{aligned}\begin{bmatrix} \xi \\ A \\ C \end{bmatrix} &= \Phi \begin{bmatrix} 0 \\ O(\varepsilon_G) + O(1)B(0)e^{-\chi_1(0)} \\ O(\varepsilon_L) + O(1)D(0)e^{-\chi_2(0)} \end{bmatrix} \\ &\quad + \Phi \int_0^z \Phi^{-1} \left\{ \begin{bmatrix} p & q \\ (b_1 + a_1')/a_1 & c_1/a_1 \\ c_2 & b_2 \end{bmatrix} \begin{bmatrix} Be^{-\chi_1} \\ De^{-\chi_2} \end{bmatrix} + \begin{bmatrix} 0 \\ -r_1/a_1 \\ -r_2 \end{bmatrix} \right\} ds\end{aligned} \tag{2.23}$$

where we have chosen $\Phi(0) = I$, the 3×3 identity.

From the boundary conditions at $z = 1$, we obtain

$$B(1) = O(\varepsilon_G), \qquad D(1) = O(\varepsilon_L), \tag{2.24}$$

so that, on integrating the two equations (2.19) and (2.20), we obtain

$$B(z) = O(\varepsilon_G) + \gamma_B \int_z^1 a_1^{-1} \gamma_B^{-1} \{b_1 A + c_1 C + d_1 \xi + c_1 De^{-\chi_2} - r_1\} e^{\chi_1} ds \tag{2.25}$$

and

$$D(z) = O(\varepsilon_L) + \gamma_D \int_z^1 \gamma_D^{-1} \{c_2 A + c_2 Be^{-\chi_1} + b_2 C + d_1 \xi - r_2\} e^{\chi_2} ds, \tag{2.26}$$

where γ_B and γ_D are integrating factors, with

$$\gamma_B = \exp\left\{\int_z^1 (b_1 + a_1')/a_1 \, ds\right\}, \qquad \gamma_D = \exp\left\{\int_x^1 b_2 \, ds\right\}.$$

We note at this point that, under the assumptions about U, X, and Y, the solutions of Problem B, γ_B, γ_D and the entries of Φ and Φ^{-1} are all functions that are continuous on $[0, 1]$ and $O(1)$ there as $\varepsilon_G, \varepsilon_L \to 0$.

Thus, the full boundary-value problem (2.2)–(2.9) for (ξ,η,ζ) has been converted to the equivalent system of integral equations (2.23), (2.25), (2.26) for the five functions ξ, A, B, C and D. Clearly, solution of these will give the solution (ξ,η,ζ) to the original problem by means of (2.11) and (2.12). Note that the above conversion tacitly assumes that $a_1(z)\neq 0$ on $[0,1]$—this is essential.

Consider now the properties of these integral equations. For each $\varepsilon_G,\varepsilon_L>0$, their right-hand sides constitute a mapping, T, from the set of continuous 5-tuplets (ξ,A,B,C,D) into itself. We introduce a norm on this set, defined by

$$\|(\xi,A,B,C,D)\|=\max\{\|\xi\|,\|A\|,\|B\|,\|C\|,\|D\|\}, \tag{2.27}$$

where $\|\xi\|,\|A\|,\|C\|$ are the usual supremum norms for functions continuous on $[0,1]$, while

$$\|B\|=\max_{z\in[0,1]}\left|B(z)e^{-\chi_1(z)}\right|,\qquad \|D\|=\max_{z\in[0,1]}\left|D(z)e^{-\chi_2(z)}\right|. \tag{2.28}$$

In terms of the norm (2.27), the ball

$$\|(\xi,A,B,C,D)\|\leq m\max\{\varepsilon_G,\varepsilon_L\} \tag{2.29}$$

is a complete metric space, for given positive m,ε_G and ε_L.

We consider the action of T on such a ball. Clearly, T is bounded on such a ball in terms of the norm (2.27). Moreover, on any such ball, provided $U(z)>0$ on $[0,1]$, $a_1(z)>0$ there for $\varepsilon_G,\varepsilon_L$ small enough, so that the exponentials in (2.25), (2.26) and the structure of r_1,r_2 allow us to verify that these components of T are contractive in terms of (2.27), with contraction parameters $O(\max\{\varepsilon_G,\varepsilon_L\})$ as $\varepsilon_G,\varepsilon_L\to 0$. While (2.23) is only bounded in terms of (2.27), successive iteration of T reveals that, for $\varepsilon_G,\varepsilon_L$ small enough, T^2 is contractive in terms of this norm, with contraction parameter $O(\max\{\varepsilon_G,\varepsilon_L\})$ as $\varepsilon_G,\varepsilon_L\to 0$.

Similar arguments also show that T^2 maps a ball (2.29) into itself, for some $m>0$ and $\varepsilon_G,\varepsilon_L$ small enough.

We thus arrive at our first basic result.

LEMMA 1. *Let Problem* B *have a solution* (U,X,Y) *that is twice continuously differentiable on* $[0,1]$ *and for which* $U(z)>0$ *there. Then, for* $\varepsilon_G,\varepsilon_L$ *sufficiently small, there is an* $m>0$ *(which is independent of* $\varepsilon_G,\varepsilon_L$*) such that the map* T *has a unique fixed point* (ξ,A,B,C,D) *in the ball* (2.29).

Proof. By the arguments above, T^2 is a contraction on such a ball, and maps it into itself, so application of the contraction mapping theorem [5, p. 26] yields the existence of a unique fixed point of T^2 in this ball. This then implies that T has such a fixed point, that coincides with that of T^2. □

We may now use Lemma 1 to obtain our first main theorem.

THEOREM 1. *Let Problem* B *have a solution* (U,X,Y) *that satisfies the hypotheses of Lemma* 1. *Then, for* $\varepsilon_G,\varepsilon_L$ *sufficiently small, Problem* A *has a unique solution* (u,x,y) *that satisfies the estimates*

$$(u-U,x-X,y-Y)=O(\max\{\varepsilon_G,\varepsilon_L\}) \tag{2.30}$$

and

$$u'-U'=O(\max\{\varepsilon_G,\varepsilon_L\}), \tag{2.31}$$

$$x'-X'=O\left(\exp\left\{-\varepsilon_G^{-1}\int_z^1 u(s)\,ds\right\}\right)+O(\max\{\varepsilon_G,\varepsilon_L\}), \tag{2.32}$$

$$y'-Y'=O\left(\exp\left\{-\varepsilon_L^{-1}(1-z)\right\}\right)+O(\max\{\varepsilon_G,\varepsilon_L\}) \tag{2.33}$$

as $\varepsilon_G,\varepsilon_L\to 0$, *uniformly with respect to* z *on* $[0,1]$.

Proof. The conclusions of Lemma 1 imply the existence of a unique solution (ξ,η,ζ) to the problem (2.2)–(2.9), and the estimates (2.30)–(2.33) follow directly from the integral equations for (ξ,A,B,C,D) and the defining relations (2.11) and (2.12). □

3. Existence for Problem B. We now turn to the question of the existence of solutions (U,X,Y) of Problem B that satisfy the requirements of Theorem 1. Clearly, since Problem B is an initial-value problem, the standard method of successive approximations will yield the existence of a unique solution in a suitably small neighbourhood of $x=0$. However, whether or not this solution may be continued throughout all of $[0,1]$, maintaining the desired properties is a question not so easily answered. We are thus interested in demonstrating the existence of such a solution. First, however, we establish two basic properties of such solutions as *do* exist.

LEMMA 2. *The solution* (U,X,Y) *of Problem* B, *when it exists and is bounded on* $[0,1]$, *is unique on* $[0,1]$.

Proof. We may regard Problem B as a problem for X and Y—once they are determined, U is uniquely determined via (1.15). The right-hand sides of the equations (1.13), (1.16) are clearly Lipschitzian in X,Y for bounded X,Y and thus, uniqueness follows by a standard argument. □

LEMMA 3. *Let* (U,X,Y) *be a solution of Problem* B *that is continuously differentiable on* $[0,1]$ *and such that* $X(z)<x_0^{-1}$ *for all* $z\in[0,1]$. *Then* U,X,Y *are all positive there, with the exception of* $Y(0)=0$.

Proof. Suppose (U,X,Y) is such a solution, and X has a zero at z_0 in $0<z_0\le 1$. Since $X(0)=1$, there is a first zero, which we may take as z_0, and which is such that

$$X(z_0)=0, \qquad X'(z_0)\le 0, \tag{3.1}$$

so that, from (1.16),

$$Y(z_0)\le 0. \tag{3.2}$$

However, from (1.13),

$$Y(z)=e^{-\sigma z}\int_0^z \sigma_L\mu^{-1}\beta(s)e^{\sigma s}X(s)\,ds, \tag{3.3}$$

where $\sigma=D+\sigma_L\mu^{-1}$; and since $X(z)\ge 0$ on $[0,z_0]$,

$$Y(z_0)>0, \tag{3.4}$$

which contradicts the above. Thus, $X(z)$ never vanishes on $[0,1]$, and consequently, neither does $Y(z)$. We may then deduce that $U(z)$ is nonvanishing on $[0,1]$ from (1.15). □

We now return to the question of the existence of a solution to Problem B. Standard results (see for example, [2, Chap. 2, §1]), tell us that this problem has a solution in some neighbourhood of $z=0$. Moreover, there exists an interval $[0,z_0]$ say, in which, by the above results,

$$0\le X(z)<x_0^{-1}, \qquad 0\le Y(z)<k, \tag{3.5}$$

for some positive constant k. Applying these bounds to the equations (1.13), (1.16), we see that, on $[0,z_0]$, we have

$$X'(z)<\sigma_G(1-x_0)^{-1}\mu^{-1}(1-x_0X)^2Y, \tag{3.6}$$

$$Y'(z)<\sigma_L X \tag{3.7}$$

for any $D\geq 0$, while

$$X(0)=1, \qquad Y(0)=0. \tag{3.8}$$

The above inequalities then give

$$Y(z)<\sigma_L x_0^{-1} z \tag{3.9}$$

and

$$X(z)<x_0^{-1}\left[1-(1-x_0)/\left(1+\sigma_G\sigma_L z^2/2\mu\right)\right], \tag{3.10}$$

which allow us to choose $z_0>1$ provided we choose $k>\sigma_L x_0^{-1}$. Thus, our solution of Problem B extends throughout all of $[0,1]$; moreover, (3.5) holds there. Thus by combining these and the results of Lemmas 2 and 3, we have our main result for this section:

THEOREM 2. *Problem* B *has a unique solution* (U,X,Y) *that is twice continuously differentiable on* $[0,1]$, *and which satisfies the bounds*

$$0<X(z)<x_0^{-1}, \tag{3.11}$$

$$Y(z)>0, \tag{3.12}$$

and

$$U(x)>0 \tag{3.13}$$

there, for all considered values of the parameters $\alpha,\sigma_L,\sigma_G,\mu,x_0$ *and* D.

Remarks. Theorem 2 now establishes the results of Theorem 1, and in particular, the equivalence of Problems A and B as $\varepsilon_G,\varepsilon_L\to 0$. Moreover, we see that, in the limit as $\varepsilon_G,\varepsilon_L\to 0$, the solution (u,x,y) of Problem A also satisfies the bounds (3.11)–(3.13).

4. Discussion. The findings of this paper should be regarded from a number of viewpoints, and with different aims in mind. Firstly, we may view the results of Theorems 1 and 2 as basic existence-uniqueness theory for a nonlinear singular perturbation problem that has physical significance and some peculiarities of its own. In particular, the incorporation of the equation for u, the bubble velocity, as well as the two distinct parameters ε_G and ε_L, plus the Neumann type boundary conditions at $z=1$ all serve to distinguish this problem from the second-order systems considered in [1] and [4]. While the techniques used by these (and other) investigators may have proved adaptable to the case at hand, it seemed more straightforward to exploit the structure of Problem A and to use the constructive method of §2 to obtain the estimates of Theorem 1, that are sharper than those of [1] or [4]. The generalization of this method to more involved systems is at present being considered by this author.

Secondly, we must consider the application of these results. In the first instance, Theorem 1 assures us of the existence of a solution to Problem A as $\varepsilon_G,\varepsilon_L\to 0$—a matter of some conjecture when viewed in the light of the numerical calculations of [3]. Moreover, the applicability of Problem B as a useful reflector of Problem A for small $\varepsilon_G,\varepsilon_L$ is a result of considerable value. Being a straightforward initial-value problem, Problem B readily lends itself to solution by standard numerical procedures. Even though the boundary layers at $z=1$ in Problem A are of "higher order" in the sense that $\varepsilon_G x''$, $\varepsilon_L y''$ are $O(1)$ there, and not $O(\varepsilon_G^{-1})$, $O(\varepsilon_L^{-1})$ respectively (as would be the case with Dirichlet boundary conditions), they still have a disruptive effect on numerical calculations in which they occur. Hence the value of Theorem 1.

While the proof of Theorem 1 is constructive, and provides an iterative scheme by which we may generate approximations to the functions ξ, η and ζ, this technique is obviously too unwieldly to be considered for practical usage. What this proof *does* do is to provide a sound theoretical basis upon which constructions employing heuristic and purely formal argument may be founded—for example, the method of matched asymptotic expansions. Note, however, that awkward as it is, the iterative scheme of Theorem 1 employs functions that are displayed explicitly, except for the entries of the matrix Φ, which remain implicit.

Finally it is of some interest whether solutions of Problem B exhibit the structural features expected in Problem A from experimental and numerical studies. While the results of [3] and [6] do not extend to the case $\varepsilon_G, \varepsilon_L \to 0$, we may take a lead from these studies, and expect Problem B to generate solutions (U, X, Y) in which X, U exhibit minima on $(0, 1)$, while Y exhibits a maximum there. While a detailed analysis of Problem B is beyond the scope of this paper, we may easily show, for example, that a necessary condition for $X(z)$ to have a minimum on $(0, 1)$ is that

$$D < \alpha,$$

which is clearly a balance between the dimensions of the column and the reactive properties of the fluid involved. Similar inequalities may be obtained for U and Y, by elementary arguments. However, the lack of an explicit solution of Problem B seems to limit the possibility of a sufficient conditon for such structure.

Acknowledgment. The author wishes to thank the referees for comments that were of considerable value in preparing this work, and Professor J. J. Mahony of the University of Western Australia for a number of personal conversations in which some of the ideas fundamental to the methods of this paper were discussed.

REFERENCES

[1] K. W. CHANG, *Singular perturbations of a boundary value problem for a vector second order differential equation*, SIAM J. Appl. Math., 30 (1976), pp. 42–54.

[2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] W.-D. DECKWER, *Non-isobaric bubble columns with variable gas velocity*, Chem. Engrg. Sci., 31 (1976), pp. 309–317.

[4] W. G. KELLEY, *A nonlinear singular perturbation problem for second order systems*, this Journal, 10 (1979), pp. 32–37.

[5] L. LIUSTERNIK AND V. SOBOLEV, *Elements of Functional Analysis*, Hindu Publishing Corp., Delhi, 1974.

[6] S. SHIOYA, N. D. P. DANG AND I. J. DUNN, *Bubble column fermenter modeling: a comparison for pressure effects*, Chem. Engrg. Sci., 33 (1978), pp. 1025–1030.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0004 $01.25/0

A GLOBAL EXISTENCE AND UNIQUENESS THEOREM FOR A RICCATI EQUATION*

J. COLBY KEGLEY†

Abstract. It is proved that a Riccati differential equation of a particular form has a unique solution satisfying the conditions that it is to exist for large values of the independent variable t and to have its graph stay above a certain line for large t. It is then proved that the solution exists for all t. Two forms of the solution are developed in terms of the confluent hypergeometric functions. An application of these results is made to an asymptotic stochastic analysis of a noisy duel problem.

1. Introduction. This paper investigates a particular form (1) in §2 of a Riccati equation that is quadratic in the independent variable t. The approach to the problem of the existence of a global solution is not from the usual initial-value standpoint, but is based on a desired feature of the solution for large t which is given by properties (i) and (ii) of the theorem in §2. The form (1) of the equation makes it easy to draw rough sketches of how the solutions behave depending on where the initial point is selected. In particular, it becomes plausible that there exists a solution defined for all t that satisfies properties (i) and (ii), but it is by no means clear that there is only one such solution. That this is the case indicates that this distinguished solution is extremely unstable. Indeed, one of the implications of Lemma 5 in §3 is that every other solution diverges from the distinguished solution as $t\to\infty$.

Our investigation is motivated by the approach used in [3] and [6] to analyze the equal-accuracy noisy duel problem for two players having finite unequal units of ammunition. This approach leads to asymptotic distributions of normalized times of first fire for the two players. The hazard rates for these distributions are expressed in terms of a solution to a Riccati equation of the form (1), and the distributions themselves are expressed in terms of a solution to a related Hermite equation.

A brief outline of these connections is given in §4. The reader may find it helpful to read that section in conjunction with the statement of the theorem to understand the reason for deriving the various properties of the distinguished solution.

2. Statement of the theorem. The principal conclusions we desire can be stated as follows.

THEOREM. *Suppose* α *is a positive number and* ϕ_1, ϕ_2 *are linear functions* $\phi_i(t)=\beta_i t+\gamma_i$, *where* $\beta_2<\beta_1$ *and* $\beta_1>0$. *Then there is exactly one solution of the Riccati equation*

$$v'(t)=\alpha[v(t)-2\phi_1(t)][v(t)-2\phi_2(t)] \tag{1}$$

that has the following two properties: There is a number t_0 *such that*

(i) *The domain of* v *includes the interval* $[t_0,\infty)$.

(ii) $v(t)-2\phi_1(t)>0$ *for* $t\geq t_0$.

Moreover, this solution has the additional properties:

(iii) *The domain of* v *is* $(-\infty,\infty)$.

(iv) $v(t)-2\phi_1(t)>0$ *for all* t.

(v) $\int^{\infty}[v(t)-2\phi_1(t)]\,dt=\int^{\infty}[v(t)-2\phi_2(t)]\,dt=\infty$.

(vi) $v(t)-2\phi_1(t)\to 0$ *as* $t\to\infty$.

(vii) $v'(t)\to 2\beta_1$ *as* $t\to\infty$.

---

* Received by the editors August 4, 1980, and in revised form December 2, 1981. This work was partially supported by the U.S. Air Force Office of Scientific Research under grant 78-3518.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

If, in addition, $\beta_2<0$, *then the following hold*:

(viii) *If* t_0 *is a number, then the conditions*

$$v(t)-2\phi_2(t)>0 \quad \text{for } t>t_0$$

and

$$v(t_0)-2\phi_2(t_0)=0$$

hold exactly when

$$\phi_1(t_0)-\phi_2(t_0)=z_0\sqrt{(\beta_1-\beta_2)/2\alpha}\,,$$

where z_0 *is the real zero of the Weber parabolic cylinder function* D_{p+1} *with* $p=\beta_1/(\beta_2-\beta_1)$.

(ix) *With* t_0 *as in* (viii) *and, for* $i=1,2$, *we define* $f_i(t)=x_i(t)\bar{\Phi}_i(t)$, *where* $x_i(t)=\alpha[v(t)-2\phi_i(t)]$ *and* $\bar{\Phi}_i(t)=\exp(-\int_{t_0}^{t}x_i(\tau)\,d\tau)$, *we have*:

(ixa) $\bar{\Phi}_1(t)<\bar{\Phi}_2(t)$ *when* $t_0<t<-t_0-2\eta$ *and* $\bar{\Phi}_1(t)>\bar{\Phi}_2(t)$ *when* $t>-t_0-2\eta$, *where* $t=-\eta$ *is the solution of* $x_1(t)=x_2(t)$;

(ixb) f_1 *is decreasing and positive on* $(-\infty,\infty)$;

(ixc) f_2 *is positive on* (t_0,∞) *and has a maximum value that occurs at a number* $t_1>-\eta>t_0$;

(ixd) $\int_{t_0}^{\infty}tf_2(t)\,dt<\infty$;

(ixe) $\int_{t_0}^{\infty}tf_1(t)\,dt=\infty$.

In the process of proving this theorem, two forms of the solutions are developed.

(A)
$$v(t)=2\phi_1(t)-\frac{\delta}{\alpha}\frac{\psi'(s)}{\psi(s)},$$

where

$$s=\delta(t+\eta), \quad \delta=\sqrt{\alpha(\beta_1-\beta_2)}\,, \quad \eta=\frac{\gamma_1-\gamma_2}{\beta_1-\beta_2},$$

and

$$\psi(s)=\zeta y_0(s)+y_1(s),$$

with:

$$\zeta=-\frac{2\Gamma(a+\frac{1}{2})}{\Gamma(a)}, \qquad a=\frac{\beta_1}{2(\beta_1-\beta_2)},$$

$$y_0(s)=s\,{}_1F_1\left(a+\frac{1}{2},\frac{3}{2};s^2\right),$$

$$y_1(s)={}_1F_1\left(a,\frac{1}{2};s^2\right),$$

and ${}_1F_1$ denotes the confluent hypergeometric function

$${}_1F_1(a,b;z)=\frac{\Gamma(b)}{\Gamma(a)}\sum_{n=0}^{\infty}\frac{\Gamma(a+n)}{\Gamma(b+n)}\frac{z^n}{n!}.$$

(B)
$$v(t)=2\phi_2(t)+\frac{\delta\sqrt{2}}{\alpha}\frac{D_{p+1}(z)}{D_p(z)},$$

where

$$z=\delta\sqrt{2}\,(t+\eta), \qquad p=\frac{\beta_1}{\beta_2-\beta_1},$$

and D_p is the Weber parabolic cylinder function (cf. [7]).

$$D_p(z)=\Gamma\left(\frac{1}{2}\right)2^{p/2}\exp\left(-\frac{z^2}{4}\right)R_p(z),$$

with

$$R_p(z)=\frac{1}{\Gamma(1/2-p/2)}\,{}_1F_1\left(-\frac{p}{2},\frac{1}{2};\frac{z^2}{2}\right)-\frac{z\sqrt{2}}{\Gamma(-p/2)}\,{}_1F_1\left(\frac{1}{2}-\frac{p}{2},\frac{3}{2};\frac{z^2}{2}\right).$$

3. Proof of the theorem. The demonstration of the conclusions is broken down into several stages.

LEMMA 1. *A function v is a solution of* (1) *on an interval I exactly when*

$$v=\phi_1+\phi_2-\frac{1}{\alpha}\frac{x'}{x}, \tag{2}$$

where $x(t)\neq 0$ for t in I and is a solution of

$$x''(t)-q(t)x(t)=0 \tag{3}$$

with

$$q=\alpha^2(\phi_1-\phi_2)^2+\alpha(\beta_1+\beta_2).$$

Proof. We rewrite (1) in the form

$$v'+2Av+Bv^2-C=0,$$

where

$$A=\alpha(\phi_1+\phi_2), \quad B=-\alpha, \quad C=4\alpha\phi_1\phi_2.$$

We then apply the result (Reid [5, p. 11]) that v is a solution of (1) on an interval I if, and only if, $v=u/x$, where $x(t)\neq 0$ on I and the pair (x,u) is a solution on I of the linear system

$$x'=Ax+Bu, \qquad u'=Cx-Au. \tag{4}$$

But this system is equivalent to equation (3), as can be seen through the connection $u=(x'-Ax)/B$. Calculating $v=u/x$ then gives the form (2).

In order to transform equation (3) into more comprehensible forms, first we make a change of independent variable.

LEMMA 2. *The general solution of equation* (1) *is*

$$v(t)=\phi_1(t)+\phi_2(t)-\frac{\delta}{\alpha}\frac{w'(s)}{w(s)},$$

where

$$s=\delta(t+\eta), \quad \delta=\sqrt{\alpha(\beta_1-\beta_2)}, \quad \eta=\frac{\gamma_1-\gamma_2}{\beta_1-\beta_2},$$

and w is a nonvanishing solution of the Weber equation

$$w''(s)+(\varepsilon-s^2)w(s)=0 \tag{5}$$

with $\varepsilon=(\beta_2+\beta_1)/(\beta_2-\beta_1)$.

While the form (5) is simpler than the form (3), it is not easy to see when its solutions are nonvanishing; the equation has an oscillatory interval about $s=0$ if $\varepsilon>0$. However, we can make the change of dependent variable $y(s)=\exp(s^2/2)w(s)$, which transforms (5) into a Hermite equation and clearly preserves the nonvanishing of solutions. In fact, going through the calculations gives the following result.

LEMMA 3. *The general solution of* (1) *is*

$$v(t)=2\phi_1(t)-\frac{\delta}{\alpha}\frac{y'(s)}{y(s)}, \tag{6}$$

where y is a nonvanishing solution of the Hermite equation

$$y''(s)-2sy'(s)-4ay(s)=0 \tag{7}$$

with

$$a=\frac{\beta_1}{2(\beta_1-\beta_2)}.$$

Before continuing, we point out that the procedure of transforming an equation of the form (3), where q is quadratic, first into the form (5) and then into the form (7) is well known. It is used, for example, in solving the time-independent Schrödinger equation for a harmonic oscillator.

Now, the general solution of (7) can be expressed in terms of the confluent hypergeometric functions. In fact, we have the following result, which may be verified by direct calculation or by referring to Slater [7, p. 100].

LEMMA 4. *Let y_0 and y_1 denote the solutions of* (7) *that satisfy the initial conditions*

$$y_0(0)=0, \qquad y_0'(0)=1,$$
$$y_1(0)=1, \qquad y_1'(0)=0.$$

Then the functions y_0 and y_1 are given by

$$y_0(s)=s\,{}_1F_1\left(a+\frac{1}{2},\frac{3}{2};s^2\right), \tag{8}$$

$$y_1(s)={}_1F_1\left(a,\frac{1}{2};s^2\right). \tag{9}$$

We now focus our attention on property (ii) in the statement of the theorem. The next result shows that, up to a multiplicative constant, there is only one solution of (7) which, when substituted in (6), can possibly work.

LEMMA 5. *Every nontrivial solution $y=c_0y_0+c_1y_1$ of* (7) *has the property that $y'(s)/y(s)\to\infty$ as $s\to\infty$ unless the constants c_0 and c_1 satisfy the relation*

$$c_0\Gamma(a)+2c_1\Gamma\left(a+\frac{1}{2}\right)=0.$$

Proof. We apply two results about confluent hypergeometric functions given in Slater [7]. We have the derivative relation (cf. [7, p. 15])

$$\frac{d}{dz}\,{}_1F_1(a,b;z)=\frac{a}{b}\,{}_1F_1(a+1,b+1;z),$$

and the asymptotic expansion as $z\to\infty$ (cf. [7, p. 60])

$${}_1F_1(a,b;z)=\frac{\Gamma(b)}{\Gamma(a)}\exp(z)z^{a-b}(1+O(z^{-1})).$$

If we now set $y=c_0y_0+c_1y_1$ where $c_0^2+c_1^2>0$ and apply these identities, then after some simplification we obtain the results that as $s\to\infty$,

$$y'(s)=\frac{\Gamma\left(\frac{1}{2}\right)\exp(s^2)s^{2a}}{\Gamma(a)\Gamma\left(a+\frac{1}{2}\right)}\left\{\frac{\Gamma(a)}{2}c_0s^{-2}\left[1+O(s^{-2})\right]+c_0\Gamma(a)\left[1+O(s^{-2})\right]\right.$$

$$\left.+2c_1\Gamma\left(a+\frac{1}{2}\right)\left[1+O(s^{-2})\right]\right\},$$

$$y(s)=\frac{\Gamma\left(\frac{1}{2}\right)\exp(s^2)s^{2a-1}}{2\Gamma(a)\Gamma\left(a+\frac{1}{2}\right)}\left\{c_0\Gamma(a)\left[1+O(s^{-2})\right]+2c_1\Gamma\left(a+\frac{1}{2}\right)\left[1+O(s^{-2})\right]\right\}.$$

Therefore, as $s\to\infty$ we have $y'(s)/2sy(s)\to 1$, which implies $y'(s)/y(s)\to\infty$, unless the constants c_0 and c_1 make this form indeterminate. But this is precisely when $c_0\Gamma(a)+2c_1\Gamma(a+\frac{1}{2})=0$.

Now recall that the variables s and t are related by $s=\delta(t+\eta)$, where $\delta>0$, so the conditions $s\to\infty$ and $t\to\infty$ are equivalent. Then the form (6) for $v(t)$ shows that a solution that exists for large t will have $v(t)-2\phi_1(t)\to-\infty$ as $t\to\infty$ unless c_0 and c_1 satisfy the relation stated in Lemma 5. Furthermore, the solution y of (7) enters into the form (6) only through the ratio y'/y, so one of the constants c_0 and c_1 may be chosen arbitrarily. For convenience, we take $c_1=1$ and then $c_0=-2\Gamma(a+\frac{1}{2})/\Gamma(a)$. We summarize what we have obtained so far as follows.

LEMMA 6. *A necessary condition for a solution v of equation* (1) *to have properties* (i) *and* (ii) *of the theorem is that*

$$v(t)=2\phi_1(t)-\frac{\delta}{\alpha}\frac{\psi'(s)}{\psi(s)}, \tag{10}$$

where $\psi(s)=\zeta y_0(s)+y_1(s)$ with $\zeta=-2\Gamma(a+\frac{1}{2})/\Gamma(a)$ and y_0,y_1 are defined by formulas (8) *and* (9), *respectively.*

Notice that (10) is of the form (A) of the solution v that is given in the remarks following the theorem. We now proceed to show that the particular solution ψ of (7) forces the corresponding solution v of (1) to have properties (i) through (ix) of the theorem.

LEMMA 7. *The solution $\psi=\zeta y_0+y_1$ of* (7) *satisfies the inequalities $\psi(s)>0$ and $\psi'(s)<0$ for all s.*

Proof. The major theoretical tool we need to prove this result is stated in the Appendix. In order to apply that theorem to our problem, we put equation (7) in self-adjoint form by multiplying both sides by $\exp(-s^2)$. The result is the equivalent equation

$$(ry')'-py=0,$$

where $r(s)=\exp(-s^2)$ and $p(s)=4ar(s)$. Since r and p are continuous with $r(s)>0$ and $p(s)>0$ for all s, we can conclude that if $y=\theta y_0+y_1$ is the solution of (7) with $\theta=\lim_{s\to\infty}-y_1(s)/y_0(s)$, then $y(s)>0$ and $y'(s)<0$ for all s. We now show that $\theta=\zeta$.

To do this, we proceed as in the proof of Lemma 5. Using the definition of y_0 and y_1, and the asymptotic expansion of the confluent hypergeometric functions again, we obtain, as $s \to \infty$,

$$y_1(s) = \frac{\Gamma(\frac{1}{2})}{\Gamma(a)} \exp(s^2) s^{2a-1} (1 + O(s^{-2}))$$

and

$$y_0(s) = \frac{\Gamma(\frac{1}{2})}{2\Gamma(a+\frac{1}{2})} \exp(s^2) s^{2a-1} (1 + O(s^{-2})).$$

This makes it clear that $\theta = -2\Gamma(a+\frac{1}{2})/\Gamma(a) = \zeta$.

By referring to the form (10) and applying the result of Lemma 7, we immediately have:

COROLLARY. *The solution* v *of* (1) *defined by* (10) *satisfies properties* (iii) *and* (iv) *and, a fortiori, satisfies properties* (i) *and* (ii).

In order to tackle properties (v) through (viii), we develop the second form (B) of the solution v.

LEMMA 8. *The solution* $\psi = \zeta y_0 + y_1$ *of* (7) *can be written in the form*

$$\psi(s) = \frac{2^a \Gamma(a+\frac{1}{2})}{\Gamma(\frac{1}{2})} \exp\left(\frac{z^2}{4}\right) D_p(z), \tag{11}$$

where $z = s\sqrt{2} = \delta\sqrt{2}(t+\eta)$ *and* D_p *is the Weber parabolic cylinder function with* $p = -2a = \beta_1/(\beta_2 - \beta_1)$.

Proof. The result follows by using the definition of D_p and simplifying the right-hand side of (11).

LEMMA 9. *The solution* v *of* (1) *defined by* (10) *can be written in the form*

$$v(t) = 2\phi_2(t) + \frac{\delta\sqrt{2}}{\alpha} \frac{D_{p+1}(z)}{D_p(z)}. \tag{12}$$

Proof. To obtain this form, first we use formula (11) for ψ and calculate ψ'/ψ. Keeping in mind that $z = s\sqrt{2}$, we obtain

$$\frac{\psi'(s)}{\psi(s)} = s + \sqrt{2} \frac{D_p'(z)}{D_p(z)}.$$

Then we use the identity (cf. [4, p. x])

$$D_p'(z) = \left(\frac{z}{2}\right) D_p(z) - D_{p+1}(z). \tag{13}$$

The result is that

$$\frac{\psi'(s)}{\psi(s)} = 2s - \sqrt{2} \frac{D_{p+1}(z)}{D_p(z)}. \tag{14}$$

Substitution of this expression into formula (10) and use of the relation $s = \delta(t+\eta)$ give the form (12) for v, which is the form (B) that was claimed.

Properties (v), (vi) and (vii) can now be attacked by using the following asymptotic expansion of the parabolic cylinder functions (cf. [4, p. x]). As $z \to \infty$,

$$D_p(z) = \exp\left(-\frac{z^2}{4}\right) z^p \left[1 - \frac{p(p-1)}{2z^2} + O(z^{-4})\right]. \tag{15}$$

LEMMA 10. *If* $\psi=\zeta y_0+y_1$, *then* $\psi(s)\to 0^+$ *as* $s\to\infty$.

Proof. We return to formula (11) for ψ, keeping in mind the connection $z=s\sqrt{2}$ and the fact that $p=-2a<0$. Substitution of the result of (15) into (11) yields, as $s\to\infty$,

$$\psi(s)=\frac{2^a\Gamma(a+\frac{1}{2})}{\Gamma(\frac{1}{2})}z^p\left[1-\frac{p(p-1)}{2z^2}+O(z^{-4})\right], \tag{16}$$

so $\psi(s)\to 0$ as $s\to\infty$. Since $\psi(s)>0$ for all s, the conclusion follows.

COROLLARY. *The solution* v *of* (1) *defined by* (10) *satisfies property* (v).

Proof. Using the form (10) with the connection $s=\delta(t+\eta)$ shows that if we fix some t_0 and let s_0 be the corresponding value of s, then

$$\int_{t_0}^{t}(v-2\phi_1)=-\frac{1}{\alpha}\int_{s_0}^{s}\left(\frac{\psi'}{\psi}\right)=-\frac{1}{\alpha}\ln\left[\frac{\psi(s)}{\psi(s_0)}\right]. \tag{17}$$

But $s\to\infty$ as $t\to\infty$, so $\int_{t_0}^{\infty}(v-2\phi_1)=\infty$ follows immediately from Lemma 10.

For the second integral, we again fix t_0 and then apply property (iv) to an interval $[t_0,t]$. The result is

$$\int_{t_0}^{t}(v-2\phi_2)>\int_{t_0}^{t}(2\phi_1-2\phi_2)=(\beta_1-\beta_2)(t^2-t_0^2)+(\gamma_1-\gamma_2)(t-t_0),$$

which $\to\infty$ as $t\to\infty$, since $\beta_1>\beta_2$.

LEMMA 11. *If* $\psi=\zeta y_0+y_1$, *then* $\psi'(s)/\psi(s)\to 0$ *as* $s\to\infty$.

Proof. We begin by using formula (14), recalling again that $z=s\sqrt{2}$. The result is

$$\frac{\psi'(s)}{\psi(s)}=\sqrt{2}\left[z-\frac{D_{p+1}(z)}{D_p(z)}\right].$$

But $z\to\infty$ as $s\to\infty$, and if we use just the result

$$D_p(z)=\exp\left(-\frac{z^2}{4}\right)z^p(1+O(z^{-2}))$$

from formula (15), we obtain

$$z-\frac{D_{p+1}(z)}{D_p(z)}=z-\frac{z^{p+1}}{z^p}\frac{[1+O(z^{-2})]}{[1+O(z^{-2})]}=\frac{zO(z^{-2})}{1+O(z^{-2})},$$

which approaches zero as $z\to\infty$.

COROLLARY. *The solution* v *of* (1) *defined by* (10) *satisfies property* (vi).

The asymptotic expansion (15) can be used again to establish property (vii). First, we isolate the most important calculation that is involved.

LEMMA 12. *The following limit relation holds for the parabolic cylinder functions*:

$$\left(\frac{D_{p+1}}{D_p}\right)'(z)\to 1 \quad \text{as } z\to\infty.$$

Proof. After using the quotient rule to calculate the indicated derivative, we use in turn the identity (13) and its companion (cf. [4, p. x])

$$D'_{p+1}(z)=(p+1)D_p(z)-\left(\frac{z}{2}\right)D_{p+1}(z).$$

The result is

$$\left(\frac{D_{p+1}}{D_p}\right)'(z)=(p+1)+\left(\frac{D_{p+1}}{D_p}\right)^2(z)-z\left(\frac{D_{p+1}}{D_p}\right)(z). \tag{18}$$

If we apply (15) and do some reshuffling of factors, we find that

$$\left(\frac{D_{p+1}}{D_p}\right)^2(z)-z\left(\frac{D_{p+1}}{D_p}\right)(z)=\frac{[1-p(p+1)/2z^2+O(z^{-4})]}{[1-p(p-1)/2z^2+O(z^{-4})]}[-p+zO(z^{-4})],$$

which approaches $-p$ as $z\to\infty$. The conclusion then follows immediately.

COROLLARY. *The solution* v *of* (1) *defined by* (10) *satisfies property* (vii).

Proof. If we look at the form (12) of the solution, we obtain

$$v'(t)=2\phi_2'(t)+\frac{\delta\sqrt{2}}{\alpha}\left(\frac{D_{p+1}}{D_p}\right)'(z)\frac{dz}{dt}=2\beta_2+2\frac{\delta^2}{\alpha}\left(\frac{D_{p+1}}{D_p}\right)'(z),$$

since $z=\delta\sqrt{2}\,(t+\eta)$. But $z\to\infty$ as $t\to\infty$, so Lemma 12 implies that

$$v'(t)\to 2\beta_2+2\frac{\delta^2}{\alpha}\quad\text{as } t\to\infty.$$

Using the definition $\delta=\sqrt{\alpha(\beta_1-\beta_2)}$ then gives the result.

Next, we use the following result about the parabolic cylinder functions.

LEMMA 13. *If* $\beta_2<0<\beta_1$ *and* $p=\beta_1/(\beta_2-\beta_1)$, *then*:

(i) $D_p(z)>0$ *for all* z;

(ii) D_{p+1} *has exactly one real zero* z_0, *and* $D_{p+1}(z)>0$ *exactly when* $z>z_0$.

Proof. The hypotheses imply that $0<p+1<1$. Hence, the result follows immediately (cf. [1, p. 126]).

COROLLARY. *The solution* v *of equation* (1) *defined by* (10) *satisfies property* (viii).

Proof. If we apply Lemma 13 to the form (12) of the solution, we see that

$$v(t)-2\phi_2(t)>0\quad\text{for } t>t_0$$

and

$$v(t_0)-2\phi_2(t_0)=0$$

exactly when t_0 satisfies $z_0=\delta\sqrt{2}\,(t_0+\eta)$. A simple calculation using the definitions of ϕ_1, ϕ_2, δ and η then gives the result.

LEMMA 14. *The functions* $\bar{\Phi}_i$ *defined in* (ix) *are related by*

$$\bar{\Phi}_2(t)=G(t)\bar{\Phi}_1(t),\tag{19}$$

where $G(t)=\exp(s_0^2-s^2)$ *with* $s=\delta(t+\eta)$ *and* $s_0=\delta(t_0+\eta)$.

Proof. Since $x_i(t)=\alpha[v(t)-2\phi_i(t)]$, we have $x_2(t)=x_1(t)+2\delta^2(t+\eta)$, from which (19) follows easily.

COROLLARY. *The functions* $\bar{\Phi}_i$ *satisfy* (ixa).

Proof. Since t_0 satisfies (viii), we have

$$(\beta_1-\beta_2)(t_0+\eta)=\phi_1(t_0)-\phi_2(t_0)=z_0\sqrt{(\beta_1-\beta_2)/2\alpha}\,,$$

which is negative because the zero z_0 of $D_{p+1}(z)$ with $0<p+1<1$ is negative (cf. [1, p. 126]). Hence, $t_0+\eta<0$ since $\beta_1>\beta_2$, so $s_0<0$. Therefore, $G(t)>1$ exactly when $s_0<s<-s_0$, i.e., when $t_0<t<-t_0-2\eta$.

LEMMA 15. *The solution* $\psi=\zeta y_0+y_1$ *of* (7) *has* $\psi''(s)>0$ *for all* s.

Proof. A simple argument for this result follows from the observation, due to the referee, that if we temporarily denote the solution $\psi=\zeta y_0+y_1$ of (7) by ψ_a, then ψ_a' satisfies (7) with a replaced by $a+\frac{1}{2}$. Upon comparing initial values, we find that $\psi_a'(s)=(-2\Gamma(a+\frac{1}{2})/\Gamma(a))\psi_{a+1/2}(s)$. Iterating this argument gives $\psi_a''(s)=4a\psi_{a+1}(s)$, which is positive by Lemma 7 and the fact that $a>0$.

COROLLARY. *The function f_1 satisfies* (ixb).

Proof. Since property (iv) implies $x_1(t)>0$ for all t, it follows from the definition of f_1 that $f_1(t)>0$ for all t. To show that f_1 is decreasing, notice that form (A) of the solution to (1) implies that

$$x_1(t)=-\frac{\delta\psi'(s)}{\psi(s)}=\frac{-d}{dt}\ln\psi(s).$$

Hence, we have

$$\bar{\Phi}_1(t)=\frac{\psi(s)}{\psi(s_0)}, \tag{20}$$

where, as in Lemma 14, $s=\delta(t+\eta)$ and $s_0=\delta(t_0+\eta)$. But then $f_1(t)=-\bar{\Phi}_1'(t)=-\delta\psi'(s)/\psi(s_0)$, so $f_1'(t)=-\delta^2\psi''(s)/\psi(s_0)<0$ by Lemmas 7 and 15.

LEMMA 16. *The function f_2 satisfies* (ixc).

Proof. That $f_2(t)>0$ for $t>t_0$ and $f_2(t_0)=0$ follows from the definition of f_2 and (viii). Next, we show that $f_2(t)\to 0$ as $t\to\infty$. For, the definition of f_2 and relation (19) imply $f_2=-\bar{\Phi}_2'=-(G\bar{\Phi}_1)'$. Using (20) then gives

$$f_2(t)=\delta G(t)\bar{\Phi}_1(t)[2s-\psi'(s)/\psi(s)]. \tag{21}$$

Since $t>t_0$ corresponds to $s>s_0$, (20) and Lemma 7 imply

$$0<\bar{\Phi}_1(t)<1 \quad \text{for } t>t_0. \tag{22}$$

Also, $\psi'(s)/\psi(s)\to 0$ as $t\to\infty$ by Lemma 11. Finally, $2sG(t)\to 0$ as $t\to\infty$ by the definition of G. Hence, f_2 is a continuous function with $f_2(t)>0$ on (t_0,∞) while $f_2(t_0)=0=f_2(\infty)$, so f_2 has an absolute maximum in (t_0,∞). To facilitate the calculation of f_2', we first use (14) and (19) to rewrite (21) as

$$f_2(t)=\delta\bar{\Phi}_2(t)\frac{D_{p+1}(z)}{D_p(z)}, \tag{23}$$

with $z=\sqrt{2}\,\delta(t+\eta)$ as before. Taking the derivative of both sides of (23) with respect to t and using (18) as well as $\bar{\Phi}_2'=-f_2$, we find that

$$f_2'(t)=2\delta^2\bar{\Phi}_2(t)\left[p+1-\frac{zD_{p+1}(z)}{D_p(z)}\right].$$

Since $t>t_0$ corresponds to $z>z_0$, where z_0 is the (negative) zero of $D_{p+1}(z)$, it follows from Lemma 13 that $f_2'(t)>0$ for $z_0<z\le 0$, i.e., $t_0<t\le -\eta$. Hence, the maximum of f_2 occurs at some $t_1>-\eta$.

In order to deal with (ixd) and (ixe), we use $f_i=-\bar{\Phi}_i'$ and integration by parts to get

$$\int_{t_0}^{t}\tau f_i(\tau)\,d\tau=t_0-t\bar{\Phi}_i(t)+\int_{t_0}^{t}\bar{\Phi}_i(\tau)\,d\tau. \tag{24}$$

Then (19) and (22) easily yield

LEMMA 17. *The function f_2 satisfies* (ixd).

Finally, (20) and the asymptotic expansion (16) for ψ show that $\bar{\Phi}_1(t)$ behaves like t^p as $t\to\infty$, where $p=\beta_1/(\beta_2-\beta_1)$. Since $\beta_2<0<\beta_1$ implies that $-1<p<0$, it follows readily from (24) that

LEMMA 18. *The function f_1 satisfies* (ixe).

4. An application to a noisy duel problem. In [3] and [6] appears a dynamic programming approach to the m vs. n equal-accuracy noisy duel problem, where the positive integers $m<n$ represent the units of ammunition the two players have. The approach begins by allowing either of the players to fire a unit of ammunition only at times corresponding to points of a discrete grid of the interval [0, 1], which is interpreted as the interval of probabilities of either player destroying the other if a unit is fired. This produces a finite sequence of simultaneous games whose 2×2 payoff matrices are determined by proceeding backwards inductively from the game where the probability of destruction is unity.

Attention is focussed on an interval of grid points at which the players have no pure strategy and which surrounds the critical probability $1/(m+n)$. It is found, under suitable hypotheses suggested by computer implementation of the above approach, that the value of the game in this interval of grid points satisfies a difference equation. Dividing both sides of this equation by an appropriate normalization factor and letting the mesh of the grid on [0, 1] approach zero leads to a normalized value $v=v_{m,n}$ of the game that satisfies a Riccati equation of the form (1) on the interval $[-1,\infty)$, with

$$\alpha=(m+n)^2\frac{m+n-2}{n-m},$$

$$\beta_1=\frac{mc_{m,n}^2}{m+n-1},\qquad \beta_2=\frac{-n}{m}\beta_1,$$

$$\gamma_1=\frac{(m+n-1)}{2(m+n)}\cdot\frac{c_{m,n}}{c_{m,n-1}}\cdot v_{m,n-1}(-1)\quad \text{for } 1\le m<n-1,$$

while $\gamma_1=0$ for $m=n-1$,

$$\gamma_2=\frac{m+n-1}{2(m+n)}\cdot\frac{c_{m,n}}{c_{m-1,n}}\cdot v_{m-1,n}(-1)\quad \text{for } 1<m\le n-1,$$

while $\gamma_2=0$ for $m=1$.

Here, it is known that the constants $c_{i,j}$ are positive for $i<j$, but analytic expressions for these constants are not known a priori. However, the hypotheses $\alpha>0$ and $\beta_2<0<\beta_1$ are evidently satisfied. Also, it is established in [3] and [6] that the initial condition $v(t_0)-2\phi_2(t_0)=0$ is to hold when $t_0=-1$. But this does not seem to be enough information to attack the existence and uniqueness problem for (1) in $[-1,\infty)$.

Instead, attention is turned to the functions defined in (ix) with $t_0=-1$, which corresponds in the normalization process to the beginning of the interval surrounding probability $1/(m+n)$ in which random strategies are to be employed. The functions $\bar{\Phi}_1(t)$ and $\bar{\Phi}_2(t)$ represent, respectively, the probability that the weaker player and the stronger player has a normalized time of first fire occurring at or after t. The functions $x_i(t)$ represent the corresponding hazard rates for the cdf's $\Phi_i(t)\equiv 1-\bar{\Phi}_i(t)$ and the functions $f_i(t)$ represent their densities.

One of the facts derived in [3] and [6] is that the weaker player's hazard rate $x_1(t)$ is to be positive for $t\ge -1$. Somewhat surprisingly, the assumption that the solution of (1) exists for large t and that $x_1(t)$ be positive for large t produces not only the global existence and uniqueness result for (1) proved herein, but also some properties of the complements $\bar{\Phi}_i(t)$ of the cdf's $\Phi_i(t)$ that could not be surmised by studying the computer runs for the 2×2 games, namely:

Property (v) implies that

$$\int_{-1}^{\infty} x_i(t)\,dt=\infty\quad \text{for } i=1,2,$$

so that

$$\lim_{t\to\infty}\bar{\Phi}_i(t)=\exp\left(-\int_{-1}^{\infty}x_i(\tau)\right)d\tau=0,$$

which implies

$$\lim_{t\to\infty}\Phi_i(t)=1.$$

This says that the probability is unity that each player fires at some time in the normalized interval $-1\leq t<\infty$ during which random strategies are employed.

Property (ixa) states that the probability is greater not only for the weaker player firing before the stronger one for normalized times near $t_0=-1$, but also for the weaker player firing after the stronger one for large t.

Properties (ixb) and (ixc) combine to show that the mode of f_1 occurs at $t_0=-1$ while that of f_2 occurs at a value of t greater than that at which the two players' hazard rates are equal, which is in turn greater than -1.

Properties (ixd) and (ixe) state that the expectation of Φ_2 is finite while that of Φ_1 is infinite.

Finally, information about the constants $c_{i,j}$ and the initial values $v_{i,j}(-1)$ for $i<j$ can be obtained by means of a complicated recursive process. First, the above-mentioned fact that $v_{m,n}(-1)-2\phi_2(-1)=0$ yields, from the formulas for $\alpha,\beta_1,\beta_2,\gamma_1$, and γ_2,

$$v_{m,n}(-1)=\frac{2n}{(m+n-1)}\cdot c_{m,n}^2+\frac{(m+n-1)}{(m+n)}\cdot\frac{c_{m,n}}{c_{m-1,n}}\cdot v_{m-1,n}(-1)\quad\text{if } 1<m\leq n-1$$

and

$$v_{1,n}(-1)=2c_{1,n}^2.$$

Then, this relation coupled with the fact from [3] and [6] that $v_{m,n}(t)-2\phi_2(t)>0$ for $t>-1$ implies, by property (viii), that

$$\phi_1(-1)-\phi_2(-1)=z_{m,n}\sqrt{(\beta_1-\beta_2)/2\alpha},$$

where $z_{m,n}$ is the zero of the Weber parabolic cylinder function D_{p+1}, with $p=-m/(m+n)$. Solving this relation for $c_{m,n}$, using the fact that $c_{m,n}>0$, gives

$$c_{m,n}=-z_{m,n}A_{m,n}+B_{m,n},$$

where

$$A_{m,n}=\left[\frac{(n-m)(m+n-1)}{2(m+n)^3(m+n-2)}\right]^{1/2}$$

and

$$B_{m,n}=\frac{(m+n-1)^2}{2(m+n)^2}\left[\frac{v_{m,n-1}(-1)}{c_{m,n-1}}-\frac{v_{m-1,n}(-1)}{c_{m-1,n}}\right]\quad\text{for } 1<m<n-1,$$

while for $n>2$,

$$B_{1,n}=\frac{n}{2(n+1)}\cdot\frac{v_{1,n-1}(-1)}{c_{1,n-1}}=\frac{n}{(n+1)}c_{1,n-1},$$

and

$$B_{n-1,n}=-\frac{(n-1)}{2n-1}\cdot\frac{v_{n-2,n}(-1)}{c_{n-2,n}},$$

and lastly that

$$B_{1,2}=0.$$

Thus, we have $c_{1,2}=-z_{1,2}/\sqrt{27}$ at the beginning of the recursive chain. Next, we find that

$$c_{1,n}=-z_{1,n}A_{1,n}+B_{1,n}$$

expresses $c_{1,n}$ in terms of $c_{1,n-1}$ for $n>2$. Hence, the formula $v_{1,n}(-1)=2c_{1,n}^2$ determines these initial values in a simple recursive way. It is clear then that the values $c_{m,n}$ and $v_{m,n}(-1)$ can eventually be calculated in terms of m,n and the zeroes $z_{m,n}$, but simple formulas for those values are not apparent.

Thus, it is indeed fortunate that the analysis presented here that is germane to the noisy duel problem does not depend on specific information about the coefficients in (1) beyond the hypotheses of the theorem. That lack of information is compensated for by the condition that properties (i) and (ii) are to hold.

Appendix. The proof of Lemma 7 depends on the following results, which can be derived by straightforward modifications of the argument given in Hille [2, §9.2].

THEOREM. *Suppose r and p are continuous functions such that $r(s)>0$ and $p(s)\geq 0$ for s real. Let y_0 and y_1 be the solutions of*

$$(ry')'-py=0 \tag{A1}$$

that satisfy the initial conditions

$$y_0(0)=0, \qquad y_0'(0)=1,$$

and

$$y_1(0)=1, \qquad y_1'(0)=0.$$

Then: (a) *The limits*

$$\theta=\lim_{s\to\infty}-\frac{y_1(s)}{y_0(s)}, \qquad \mu=\lim_{s\to\infty}-\frac{y_1'(s)}{y_0'(s)}$$

exist, and $\theta\leq\mu$.

(b) *The solutions of* (A1) *passing through the point* $(0,1)$ *that have $y(s)>0$ and $y'(s)\leq 0$ for $s>0$ are precisely those solutions $y=\lambda y_0+y_1$ that have $\theta\leq\lambda\leq\mu$. Moreover, every such solution satisfies*:

(b_1) *$y(s)>0$ and $y'(s)\leq 0$ for all real s.*

(b_2) *$y'(s)<0$ over any interval on which $p(s)$ does not vanish identically.*

(c) *$\theta=\mu$ exactly when*

$$\int_0^\infty\left[r(y_0')^2+p(y_0)^2\right]=\infty,$$

a sufficient condition for which is the divergence of $\int_0^\infty 1/r$.

REFERENCES

[1] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.
[2] EINAR HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1969.
[3] J. COLBY KEGLEY, *An asymptotic stochastic view of anticipation in a noisy duel* (II), Preprint.
[4] I. YE. KIREYEVA AND K. A. KARPOV, *Tables of Weber Functions*, Vol. I, Pergamon, New York, 1961.
[5] W. T. REID, *Riccati Differential Equations*, Mathematics in Science and Engineering, vol. 86, Academic Press, New York, 1972.
[6] DAN R. ROYALTY, J. COLBY KEGLEY, H. T. DAVID AND R. W. BERGER, An *asymptotic stochastic view of anticipation in a noisy duel* (I), Preprint.
[7] LUCY SLATER, *Confluent Hypergeometric Functions*, Cambridge University Press, London, 1960.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0005 $01.25/0

COMPARISON THEOREMS FOR SECOND ORDER NONSELFADJOINT DIFFERENTIAL SYSTEMS*

E. C. TOMASTIK†

Abstract. Comparison theorems for conjugate and focal points of $(rx')'+px=0$ are given where $r(t)$ and $p(t)$ are continuous $n\times n$ matrices. No sign restrictions are made on the elements of r and p, but certain restrictions are made on the comparing equation. All results are new even in the selfadjoint case.

Comparison theorems play a very important role in the existence and location of zeros of differential equations, which in turn play an important role in boundary value problems.

Consider the two second order ordinary differential equations

$$\text{(i)} \qquad (r(t)x')'+p(t)x=0,$$

$$\text{(ii)} \qquad (R(t)y')'+P(t)y=0.$$

The classical Sturm comparison theorem states that if "$P(t)$ is larger than $p(t)$" and "$r(t)$ is larger than $R(t)$", then (ii) oscillates "faster" than (i). More precisely, if there exists a solution of (i) with $x(a)=x(b)=0$, $a<b$, then there must exist a solution $y(t)$ of (ii) such that $y(a)=y(c)=0$, where $c\in(a,b)$. Here the term "$P(t)$ is larger than $p(t)$" means that $P(t)\geq p(t)$ on $[a,b]$ and that this inequality becomes a strict inequality for some point in $[a,b]$.

Marston Morse [7] was the first person to obtain a comparison theorem for systems of equations. For this result one must assume in (i) and (ii) that $r(t),p(t),R(t),P(t)$ are all symmetric $n\times n$ matrices (the selfadjoint case) and that $x(t)$ and $y(t)$ are (column) n-vectors. Morse showed that if $r(t)$ and $R(t)$ are positive definite matrices for all $t\in[a,b]$ and if "$P(t)$ is larger than $p(t)$" and "$R(t)$ is larger than $r(t)$", then (ii) oscillates "faster" than (i). More precisely, if (i) has a vector solution $x(t)$ such that $x(a)=x(b)=0$, $a<b$, then there must exist a vector solution $y(t)$ of (ii) such that $y(a)=y(c)=0$, where $c\in(a,b)$. Morse defined the term "$P(t)$ is larger than $p(t)$" to mean that the matrix $P(t)-p(t)$ is positive semidefinite for all $t\in[a,b]$ and positive definite for at least one point in $[a,b]$. For other results and further references in this direction, consult the books of Coppel [5], Morse [8], and Reid [9].

More recently Ahmad and Lazer [1] established another comparison theorem for the two n-dimensional systems (i) and (ii). They assumed that $r(t)$ and $R(t)$ were the identity matrix and that the off-diagonal elements of $p(t)$ and $P(t)$ were nonnegative. They then obtained the same result as Morse did only using a different definition of "$P(t)$ is larger than $p(t)$". Ahmad and Lazer defined the term "$P(t)$ is larger than $p(t)$" to mean that the matrix $P(t)-p(t)$ has all nonnegative elements on $[a,b]$ and that one element is positive at one poinit on $[a,b]$. Since $P-p$ being positive definite is independent of $P-p$ having nonnegative elements, the two results of Morse and Ahmad and Lazer are independent. Ahmad and Lazer [2] then extended their results to the nonselfadjoint case by dropping the condition that $p(t)$ and $P(t)$ be symmetric matrices. Cheng [4], Keener and Travis [6], Schmitt and Smith [10], and Smith [11] have also recently established comparison theorems for the nonselfadjoint systems (i) and (ii) by also assuming that certain or all elements of $p(t)$ and $P(t)$ are nonnegative.

* Received by the editors April 9, 1981, and in revised form February 4, 1982.

† Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268.

In this paper is given a comparison theorem for systems somewhat similar to that of Ahmad and Lazer. In addition are given two comparison theorems for a "focal point" problem. Conditions given in this paper are independent of the conditions given in the Morse comparison theorem even in the selfadjoint case. Also the restrictions of Ahmad and Lazer and the others mentioned above that $r(t)$ and $R(t)$ be the identity matrix and that the off-diagonal elements of $p(t)$ must be nonnegative are all dropped. However, the result of the first theorem given here is somewhat weaker than that obtained by Ahmad and Lazer. In [12] the author gave a comparison theorem for a general nonlinear system that includes (i) in the case where $r(t)$ is the identity matrix, but in this general context the equation that (i) was compared to was required to be a scalar equation. This requirement is not made here.

It should be emphasized that all the results in this paper are new even in the selfadjoint case.

If y is a vector or matrix, $y>0$ ($y\geq 0$) shall mean that each element of y is positive (nonnegative). If y is a vector, y_i will denote its ith component. If q is a matrix, q_{ij} will denote the element in the ith row and jth column. If y is an n-vector, then $|y|=|y_1|+\cdots+|y_n|$.

It will be assumed throughout this paper that $r(t),p(t),R(t)$ and $P(t)$ are all continuous $n\times n$ matrices on $[a,b]$, that $r(t)$ and $R(t)$ are nonsingular on $[a,b]$ (to insure existence and uniqueness), and that furthermore $P(t)\geq 0$, $R^{-1}(t)\geq 0$, no row of $P(t)$ is identically zero, and all diagonal elements of $R^{-1}(t)$ are strictly positive.

The first theorem is a comparison theorem for conjugate points and requires that $r(t)$ and $R(t)$ be diagonal. The comparison theorems for focal points given later do not require this restriction.

THEOREM 1. *Suppose $r(t)$ and $R(t)$ are diagonal with $r(t)=\operatorname{diag}(r_1(t),\cdots,r_n(t))$ and $R(t)=\operatorname{diag}(R_1(t),\cdots,R_n(t))$, and suppose that $y(t)$ is a solution of*

$$(R(t)y')'+P(t)y=0, \qquad y(a)=0,$$

with $y(t)>0$ on (a,b). If $|r_i^{-1}(t)|\,|p_{ij}(t)|\leq R_i^{-1}(t)P_{ij}(t)$, $1\leq i,j\leq n$, on $[a,b]$ and if for any $i=1,\cdots,n$, there exists a $j=j(i)$, $1\leq j\leq n$, and $t_i\in[a,b]$ such that $|r_i^{-1}(t_i)|\,|p_{ij}(t_i)|<R_i^{-1}(t_i)P_{ij}(t_i)$, then

$$(r(t)x')'+p(t)x=0, \qquad x(a)=0=x(c) \tag{1}$$

has no nontrivial solution for any $c\in(a,b]$.

Proof. To begin, define the diagonal matrix $g=\operatorname{diag}(g_1,\cdots,g_n)$ by

$$g_i(t,s,c,r_i)=\begin{cases}\left(\int_a^c r_i^{-1}(u)\,du\right)^{-1}\int_t^c r_i^{-1}(u)\,du\int_a^s r_i^{-1}(u)\,du, & a\leq s\leq t\leq c,\\ \left(\int_a^c r_i^{-1}(u)\,du\right)^{-1}\int_a^t r_i^{-1}(u)\,du\int_s^c r_i^{-1}(u)\,du, & a\leq t\leq s\leq c,\end{cases}$$

$1\leq i\leq n$. Then it readily follows that $x(t)$ is a solution of (1) if and only if $x(t)$ satisfies

$$x(t)=\int_a^c g(t,s,c,t)p(s)x(s)ds.$$

Using this notation,

$$y(t)=\int_a^b g(t,s,b,R)P(s)y(s)\,ds+\left(\int_a^b R^{-1}(u)\,du\right)^{-1}\int_a^t R^{-1}(u)\,du\,y(b). \tag{2}$$

Notice that

$$y'(a)=\left(\int_a^b R^{-1}(u)\,du\right)^{-1}R^{-1}(a)\int_a^b\int_s^b R^{-1}(u)\,du\,P(s)y(s)\,ds$$
$$+\left(\int_a^b R^{-1}(u)\,du\right)^{-1}R^{-1}(a)y(b)>0$$

by the hypothesis on R, P, and y. Also notice that, since

$$y'(b)=-\left(\int_a^b R^{-1}(u)\,du\right)^{-1}R^{-1}(b)\int_a^b\int_a^s R^{-1}(u)\,du\,P(s)y(s)\,ds$$
$$+\left(\int_a^b R^{-1}(u)\,du\right)^{-1}R^{-1}(b)y(b),$$

$y_i'(b)<0$ if $y_i(b)=0$. It then follows that each of the terms $x_i(t)y_i^{-1}(t)$ is continuous on $[a,c]$. Now define $\|x_i\|=\sup(|x_i(t)|y_i^{-1}(t): t\in[a,c])$ and $\|x\|=\max(\|x_i\|: i=1,\cdots,n)$. Then for any $t\in[a,c]$

$$|x_i(t)|=\left|\sum_k\int_a^c g_i(t,s,c,r_i)p_{ik}(s)x_k(s)\,ds\right|$$
$$\leq\sum_k\int_a^c g_i(t,s,c,|r_i|)|p_{ik}(s)||x_k(s)|\,ds$$
$$=\sum_k\int_a^c g_i(t,s,c,|r_i|)|p_{ik}(s)|y_k(s)|x_k(s)|y_k^{-1}(s)\,ds$$
$$\leq\sum_k\int_a^c g_i(t,s,c,|r_i|)|p_{ik}(s)|y_k(s)\,ds\,\|x\|.$$

It follows that for $t\in(a,c)$,

$$|x_i(t)|<\sum_k\int_a^c g_i(t,s,c,R_i)P_{ik}(s)y_k(s)\,ds\,\|x\|,$$

and then

$$|x_i(t)|y_i^{-1}(t)<y_i^{-1}(t)\sum_k\int_a^c g_i(t,s,c,R_i)P_{ik}(s)y_k(s)\,ds\,\|x\|. \tag{3}$$

This last strict inequality will now be extended to $[a,c)$. This will be done by showing that

(4)

$$y_i^{-1}(t)\sum_k\int_a^c g_i(t,s,c,|r_i|)|p_{ik}(s)|y_k(s)\,ds<y_i^{-1}(t)\sum_k\int_a^c g_i(t,s,c,R_i)P_{ik}(s)y_k(s)\,ds$$

on $[a,c)$. Toward this end notice that it has already been shown that each $y_i(t)$ has a zero at $t=a$ of precisely order one. Of course, $x_i(t)$, $\Sigma_k\int_a^c g_i(t,s,c,|r_i|)|p_{ik}(s)|y_k(s)\,ds$, and $\Sigma_k\int_a^c g_i(t,s,c,R_i)P_{ik}(s)y_k(s)\,ds$ each has a zero at $t=a$ of at least order one. Taking the limit as $t\to a+$ of the left-hand side of (4) readily yields

$$(y_i'(a))^{-1}\left(\int_a^c|r_i^{-1}(u)|\,du\right)^{-1}|r_i^{-1}(a)|\sum_k\int_a^c\int_s^c|r_i^{-1}(u)|du\,|p_{ik}(s)|y_k(s)\,ds,$$

while taking the limit as $t \to a+$ of the right-hand side of (4) yields

$$(y_i'(a))^{-1}\left(\int_a^c R_i^{-1}(u)\,du\right)^{-1} R_i^{-1}(a)\sum_k \int_a^c \int_s^c R_i^{-1}(u)\,du\,P_{ik}(s)y_k(s)\,ds.$$

It is now clear that the second limit is strictly larger than the first limit. This shows that (4) and thus (3) holds on $[a,c)$.

A computation shows that

$$\frac{\partial}{\partial\beta}\int_a^\beta g(t,s,\beta,R)P(s)y(s)\,ds$$

$$= R^{-1}(\beta)\left(\int_a^\beta R^{-1}(u)\,du\right)^{-2}\int_a^t R^{-1}(u)\,du\int_a^\beta\int_a^s R^{-1}(u)\,ds\,P(s)y(s)\,ds \geq 0$$

if $\beta > a$, $t \geq a$. Thus

$$\int_a^c g(t,s,c,R)P(s)y(s)\,ds \leq \int_a^b g(t,s,b,R)P(s)y(s)\,ds,$$

since $a < c \leq b$. Then from (3)

$$|x_i(t)|y_i^{-1}(t) < y_i^{-1}(t)\sum_k \int_a^b g_i(t,s,b,R)P_{ik}(s)y_k(s)\,ds\,\|x\|$$

for $t \in [a,c)$. If $c < b$, this extends immediately to $[a,c]$ since $x_i(c) = 0$ and $\Sigma_k \int_a^b g_i(c,s,b,R_i)P_{ik}(s)y_k(s)\,ds > 0$. Now suppose that $c = b$. First assume $y_i(b) = 0$. Then taking limits as $t \to b-$ on each side of (4), in the very same way as was done for $t \to a+$, yields (3) for all $t \in [a,c]$. Since $y_i(b) \geq 0$, it follows in all these cases that

$$|x_i(t)|y_i^{-1}(t) < y_i^{-1}(t)\left[\sum_k \int_a^b g_i(t,s,b,R_i)P_{ik}(s)y_k(s)\,ds + \left(\int_a^b R_i^{-1}(u)\,du\right)^{-1}\int_a^t R_i^{-1}(u)\,du\,y_i(b)\right]\|x\|$$

for $t \in [a,c]$. In the one remaining case that $c = b$ and $y_i(b) > 0$, this last inequality holds since $x_i(c) = 0$. If we recall (2), this last inequality is just

$$|x_i(t)|y_i^{-1}(t) < y_i^{-1}(t)[y_i(t)]\|x\| = \|x\|$$

for all $t \in [a,c]$. Since $|x_i(t)|y_i^{-1}(t)$ is continuous on $[a,c]$, this implies that $\|x_i\| < \|x\|$ and thus $\|x\| < \|x\|$. This contradiction then establishes Theorem 1. □

THEOREM 2. *Suppose that $y(t)$ is a solution of*

$$(R(t)y')' + P(t)y = 0, \qquad y'(b) = 0,$$

with $y(t) > 0$ on (a,b). If no diagonal element of $r^{-1}(t)$ is ever zero on $[a,b]$ and if $\int_a^t |r_{ij}^{-1}(u)|\,du \leq \int_a^t R_{ij}^{-1}(u)\,du$ and $|p_{ij}(t)| \leq P_{ij}(t)$ on $[a,b]$ for $1 \leq i,j \leq n$ and if for any $i = 1,\cdots,n$ there exists $j = j(i)$, $1 \leq j \leq n$, and $t_i \in [a,b]$ such that $|p_{ij}(t_i)| < P_{ij}(t_i)$, then

$$(r(t)x')' + p(t)x = 0, \qquad x(c) = 0 = x'(b), \tag{5}$$

has no nontrivial solution for any $c \in [a,b)$.

Proof. To begin, define the matrix $g(t,s,a,r)$ by

$$g(t,s,a,r) = \begin{cases} \int_c^t r^{-1}(u)\,du, & c \leq t \leq s \leq b, \\ \int_c^s r^{-1}(u)\,du, & c \leq s \leq t \leq b. \end{cases}$$

Then $x(t)$ is a solution to (5) if and only if $x(t)$ is a solution of

$$x(t)=\int_c^b g(t,s,c,r)p(s)x(s)\,ds.$$

When we make note of the facts

$$y(t)=\int_a^b g(t,s,a,R)P(s)y(s)\,ds+y(a),$$

$$y'(a)=R^{-1}(a)\int_a^b P(s)y(s)\,ds>0,$$

and

$$\frac{\partial}{\partial\alpha}\int_\alpha^b g(t,s,\alpha,R)P(s)y(s)\,ds=-R^{-1}(\alpha)\int_\alpha^b P(s)y(s)\,ds\leq 0,$$

the proof follows similar lines as in the previous theorem. The proof is actually simpler since no component of

$$y(b)=\int_a^b\int_a^s R^{-1}(u)\,du\,P(s)y(s)\,ds$$

is zero. □

THEOREM 3. *Suppose that $y(t)$ is a solution of*

$$(R(t)y')'+P(t)y=0,\qquad y'(a)=0,$$

with $y(t)>0$ on (a,b). If no diagonal element of $r^{-1}(t)$ is ever zero on $[a,b]$ and if $\int_t^b|r_{ij}^{-1}(u)|\,du\leq\int_t^b R_{ij}^{-1}(u)\,du$ and $|p_{ij}(t)|\leq P_{ij}(t)$ on $[a,b]$ for $1\leq i,j\leq n$ and if for any $i=1,\cdots,n$ there exists $j=j(i)$, $1\leq j\leq n$, and $t_i\in[a,b]$ such that $|p_{ij}(t_i)|<P_{ij}(t_i)$, then

$$(r(t)x')'+p(t)x=0,\qquad x'(a)=0=x(c), \tag{6}$$

has no nontrivial solution for any $c\in(a,b]$.

Proof. Define the matrix $g(t,s,c,r)$ by

$$g(t,s,c,r)=\begin{cases}\int_t^b r^{-1}(u)\,du, & c\leq s\leq t\leq b,\\ \int_s^b r^{-1}(u)\,du, & c\leq t\leq s\leq b.\end{cases}$$

Then $x(t)$ is a solution of (6) if and only if $x(t)$ satisfies

$$x(t)=\int_c^b g(t,s,c,r)p(s)x(x)\,ds.$$

The proof then proceeds as in the previous theorem. □

REFERENCES

[1] S. AHMAD AND A. C. LAZER, *On the components of extremal solutions of second order systems*, this Journal, 8 (1977), pp. 16–23.

[2] ———, *An N-dimensional extension of the Sturm separation and comparison theory to a class of non-selfadjoint systems*, this Journal, 9 (1978), 1137–1150.

[3] S. AHMAD, *On positivity of solutions and conjugate points on nonselfadjoint systems*, to appear.

[4] S. CHENG, *Nonoscillatory solutions of $x^{(m)}=(-1)^m Q(t)x$*, Canad. Math. Bull., 22 (1979), pp. 17–21.

[5] W. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics 20, Springer-Verlag, Berlin-Heidelberg-New York, 1971.

[6] M. S. KEENER AND C. C. TRAVIS, *Sturmian theory for a class of nonselfadjoint differential systems*, Ann. Mat. Pura ed Appl., 123 (1980), pp. 247–266.

[7] M. MORSE, *A generalization of the Sturm separation and comparison theorems in n-space*, Math. Ann., 103 (1930), pp. 53–69.
[8] _____, *Variational Analysis: Critical Extremals and Sturmian Extensions*, John Wiley, New York, 1973.
[9] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
[10] K. SCHMITT AND H. L. SMITH, *Positive solutions and conjugate points for systems*, Nonlinear Anal., 2 (1978), pp. 93–105.
[11] H. L. SMITH, *A note on disconjugacy for second order systems*, Pacific J. Math., 89 (1980), pp. 447–452.
[12] E. C. TOMASTIK, *Conjugate and focal points of second order differential systems*, this Journal, 12 (1981), pp. 314–320.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0006 $01.25/0

A NONLINEAR VOLTERRA INTEGRODIFFERENTIAL EQUATION DESCRIBING THE STRETCHING OF POLYMERIC LIQUIDS*

P. MARKOWICH† AND M. RENARDY‡

Abstract. We study a model equation for the elongation of filaments or sheets of polymeric liquids under the influence of a force applied to the ends. Mathematically this equation has the form of a nonlinear Volterra integrodifferential equation with the kernel given by a finite sum of exponentials. The unknown function denotes the length of the filament or, respectively, the thickness of the sheet. We study the equation both analytically and numerically. The force is assumed to converge to zero exponentially as $t \to -\infty$ and to vanish identically after a finite time t_0. It is shown that under this condition there is a unique solution which approaches a given limit as $t \to -\infty$; moreover, the solution also has a limit as $t \to +\infty$. A numerical scheme is analyzed and convergence uniform in t is established. Particular attention is paid to the dependence of solutions on a parameter μ, which corresponds to a Newtonian contribution to the viscosity. It is proved that solutions converge uniformly in t as $\mu \to 0$, and that the convergence of the numerical scheme is also uniform in μ.

Key words. viscoelastic liquids, nonlinear Volterra integrodifferential equations, singular perturbation, numerical approximation on infinite intervals.

AMS-MOS subject classification (1980). Primary 34D05, 34D15, 45J05, 45L10, 65R20, 76A10.

1. Introduction. In this paper we consider a mathematical model describing the stretching of a filament or a sheet of a molten polymer under a prescribed force f. These two physical situations are illustrated by the diagrams in Fig. 1.

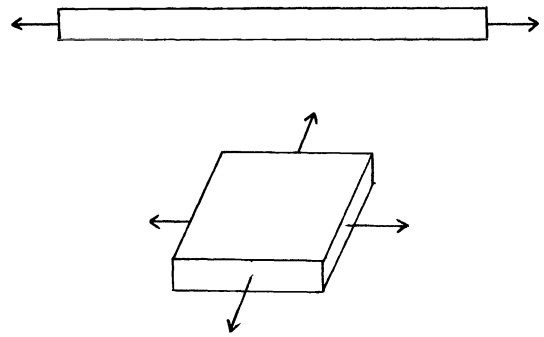


FIG. 1

Our model is based on the following physical assumptions (for more background material on the problem, we refer to Petrie [18]):

(i) The polymer satisfies the "rubberlike liquid" constitutive relation [5].

(ii) The strain and stress tensors are independent of spatial coordinates, and, in particular, inertial forces are neglected (for a model that includes inertial forces see [9]).

(iii) The molten polymer is incompressible.

* Received by the editors June 4, 1981, and in revised form November 20, 1981. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041.

† Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, A-1040 Wien, Austria. This material is based upon work supported by the National Science Foundation under grant MCS-79-27062, and by the Austrian Ministry for Science and Research.

‡ Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was supported by Deutsche Forschungsgemeinschaft.

Under these assumptions the problem is described by the equation (for a derivation see [6], [9]):

$$(1.1)\qquad \mu\dot{y}(t)+\int_{-\infty}^{t} a(t-s)\left(\frac{y^3(t)}{y^2(s)}-y(s)\right)ds=f(t)y^{\alpha}(t),\qquad -\infty<t<\infty,$$

where y denotes the length of the filament or the thickness of the sheet, respectively, μ is a nonnegative material constant modelling a Newtonian contribution to the viscosity, which may physically come from a solvent or fractions of low molecular weight, and the memory kernel a has the form

$$(1.2)\qquad a(u)=\sum_{l=1}^{N} K_l e^{-\lambda_l u}$$

with positive constants K_l, λ_l. f denotes the force acting on the ends of the filament, or $-f$ denotes the force acting on the edges of the sheet, respectively. The exponent α is 2 for the filament and $\frac{1}{2}$ for the sheet (for our mathematical analysis, we assume $0<\alpha<3$). The difference comes from geometric reasons: If f were denoting the force per unit area, α would be 1 for both cases. Due to the incompressibility, however, the area on which f is acting depends on y.

Although this has no significance to the mathematical analysis, the physical relevance of the model is limited to $f\geq 0$ for the filament and $f\leq 0$ for the sheet. If, for example, one attempted to compress the filament, then buckling rather than contraction would be observed, and this instability is not described by our equation.

A problem related to ours was investigated by Lodge, McLeod and Nohel [6]. They assume $y(t)$ is given for $t\leq 0$, it is nondecreasing (which implies but does not follow from $f\geq 0$), and $y(-\infty)=1$. They then assume $f=0$ for $t>0$ and study the elastic recovery. For a class of kernels a and functions $F(y(t),y(s))$ under the integral, which include those specified above, they prove the existence of a unique solution to the history value problem, which is nondecreasing for $t>0$ and converges to a limit $y(\infty)>1$. Their proofs rely on monotonicity arguments, and they also prove that the solutions depend monotonically on the prescribed history and the parameter μ. One of the main points in their analysis is the behavior of solutions near $\mu=0$; in this case the solutions become discontinuous at $t=0$, and they face a singular perturbation problem with a boundary layer. On the basis of these results Nevanlinna [8] used an implicit first order Euler-type discretization scheme for (1.1). He proved that this discretization preserves all the monotonicity properties, and that the global error is $O(h^{\gamma})$ for some $\gamma<1$ uniformly with respect to $\mu\in[0,\mu_0]$ and $t\in[t_0,\infty)$, $t_0>0$. An error estimate $O(h)$ uniformly in μ was shown only for $t\in[t_0,T]$ with T finite.

In our analysis, we prescribe a continuous function $f(t)$, which satisfies $\lim_{t\to-\infty} e^{-\sigma t}f(t)=0$ for some $\sigma>0$, and $f\equiv 0$ for $t\in[t_0,\infty)$. We prove that, for any such f, problem (1.1) has a unique solution $y(t)$ satisfying $\lim_{t\to-\infty} y(t)=1$. This convergence is exponential; moreover, the solution exists globally in time and converges exponentially to a constant $y(\infty)>0$ as $t\to\infty$; more precisely, we have $\lim_{t\to-\infty} e^{-\sigma t}(y(t)-1)=\lim_{t\to+\infty} e^{\sigma t}(y(t)-y(\infty))=0$. This holds for any $\mu\geq 0$. The solution depends continuously on μ in a norm stronger than the L^{∞}-norm (more specifically, in an exponentially weighted L^{∞}-norm, which incorporates the asymptotic behavior as $t\to\pm\infty$), even at $\mu=0$. No boundary layer occurs, since the solution for

$\mu=0$ has the correct asymptotic behavior as $t\to\pm\infty$. Our proofs are mainly based on the implicit function theorem and Lyapunov function arguments.

In the second part of the paper we discuss the computational solution of (1.1). Like Nevanlinna, we use a first order implicit Euler-type discretization with uniform mesh size h, after having cut the interval $[-\infty,0]$ at t_{-m}. In the convergence proof, we use a discrete analogue of exponentially weighted L^∞-spaces (infinite sequences converging exponentially on both sides). Choosing a space with an exponential weight given by $e^{(\sigma-\varepsilon)|t|}$, $0\le\varepsilon<\sigma$, we obtain an error estimate of the form $O(h)+o(e^{-\varepsilon|t-m|})$ in the norm of that space; moreover, this holds uniformly in $\mu\in[0,\infty)$ and $\varepsilon\in[0,\varepsilon_0]$, $\varepsilon_0<\sigma$. The main tool in the proof is Keller's nonlinear stability concept [3].

Our numerical results imply that the solution $y(t,\mu)$ does not differ significantly from $y(t,0)$ on $[-\infty,\infty]$ if μ is smaller than a certain fairly large number. If μ exceeds this number, then the solutions change considerably.

The paper is organized as follows: In §2 we present the analytical results, §3 concerns the discretization procedure, and the computations are reported in §4.

2. Analysis of the continuous problem.

Solutions for small forces. Let us consider (1.1), where $0<\alpha<3$, $a(u)=\sum_{i=1}^N K_i e^{-\lambda_i u}$, and $\mu>0$. This equation can be reduced to a system of ODEs in two ways. We set

$$g_i(t)=\int_{-\infty}^t K_i e^{-\lambda_i(t-s)}\frac{1}{y^2(s)}\,ds, \qquad h_i(t)=\int_{-\infty}^t K_i e^{-\lambda_i(t-s)}y(s)\,ds.$$

Then (1.1) reads

$$\begin{aligned}
\dot y &= -\frac{1}{\mu}\left(\sum_{x=1}^N (g_i y^3-h_i)-f(t)y^\alpha\right),\\
\dot g_i &= -\lambda_i g_i+\frac{K_i}{y^2},\\
\dot h_i &= -\lambda_i h_i+K_i y.
\end{aligned}\tag{2.1}$$

If we set $\gamma_i=g_i y^2$, $\delta_i=h_i/y$, we obtain

$$\begin{aligned}
\dot y &= -\frac{1}{\mu}\left(\sum_{i=1}^N (\gamma_i-\delta_i)y-f(t)y^\alpha\right),\\
\dot\gamma_i &= -\lambda_i\gamma_i+K_i-\frac{2}{\mu}\gamma_i\sum_j(\gamma_j-\delta_j)+\frac{2}{\mu}\gamma_i f(t)y^{\alpha-1},\\
\dot\delta_i &= -\lambda_i\delta_i+K_i+\frac{1}{\mu}\delta_i\sum_j(\gamma_j-\delta_j)-\frac{1}{\mu}\delta_i f(t)y^{\alpha-1}.
\end{aligned}\tag{2.2}$$

Both forms (2.1) and (2.2) will be used in the following.

Clearly, if $f=0$, then $y=1$, $g_i=h_i=K_i/\lambda_i=\gamma_i=\delta_i$ is a stationary solution.

LEMMA 2.1. *The $(2N+1)$-square matrix setting up the right-hand side of the linearization of* (2.1) (*or* (2.2)) *at the stationary solution $y=1$, $g_i=h_i=K_i/\lambda_i$ has zero as a simple eigenvalue. All other eigenvalues are real and negative.*

Proof. Clearly, (2.1) and (2.2) give the same eigenvalues. Let us consider (2.1). The linearization is set up by the following matrix:

$$A=\begin{bmatrix} -\sum_{i=1}^{N}\frac{3K_i}{\mu\lambda_i} & -\frac{1}{\mu} & -\frac{1}{\mu} & \cdots & -\frac{1}{\mu} & \frac{1}{\mu} & \frac{1}{\mu} & \cdots & \frac{1}{\mu} \\ -2K_1 & -\lambda_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ -2K_2 & 0 & -\lambda_2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ -2K_N & 0 & 0 & \cdots & -\lambda_n & 0 & 0 & \cdots & 0 \\ K_1 & 0 & 0 & \cdots & 0 & -\lambda_1 & & \cdots & 0 \\ K_2 & 0 & 0 & \cdots & 0 & 0 & -\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ K_N & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\lambda_n \end{bmatrix}$$

This yields the characteristic polynomial

$$P(\lambda)=\prod_j(-\lambda_j-\lambda)^2\left(-\sum_i\frac{3K_i}{\mu\lambda_i}-\lambda-\sum_i\frac{3K_i}{\mu(-\lambda_i-\lambda)}\right).$$

Thus N eigenvalues are given by $\lambda=-\lambda_i$; the remaining $N+1$ eigenvalues are the zeros of the last factor. Obviously one of these is zero, and it is simple. It remains to be proved that all the remaining roots have negative real parts. Consider the equation

$$-\sum_i\frac{3K_i}{\mu\lambda_i}-\lambda-\sum_i\frac{3K_i}{\mu(-\lambda_i-\lambda)}=0. \tag{2.3}$$

The left-hand side has poles at $\lambda=-\lambda_i$, and its sign is positive for $\lambda\to-\lambda_i+$ and negative for $\lambda\to-\lambda_i-$. For convenience, let the λ_i's be ordered such that $\lambda_1<\lambda_2<\cdots<\lambda_N$. It follows that there is a root in each interval $(-\lambda_i,-\lambda_{i+1})$ and another root between $-\lambda_N$ and $-\infty$. Hence all nonzero roots are real and negative. □

We want to prove the existence of solutions for small f using the implicit function theorem. The spaces in which we apply this theorem are defined in the following:

DEFINITION 2.2. Let $Y^{\sigma,n}=\{g\in C^n(\mathbb{R},\mathbb{R})|\lim_{|t|\to\infty}e^{\sigma|t|}g^{(k)}(t)=0 \text{ for } k=0,1,\cdots,n\}$. A natural norm in $Y^{\sigma,n}$ is

$$\|g\|=\sum_{k=0}^{n}\sup_{t\in\mathbb{R}}|e^{\sigma|t|}g^{(k)}(t)|.$$

Moreover, let $X^{\sigma,n}=\{f\in C^n(\mathbb{R},\mathbb{R})|\lim_{|t|\to\infty}e^{\sigma|t|}f^{(k)}(t)=0 \text{ for } k=1,\cdots,n, \exists f(\infty)$ such that $\lim_{t\to\infty}e^{\sigma t}(f(t)-f(\infty))=\lim_{t\to-\infty}e^{-\sigma t}f(t)=0\}$. A natural norm in $X^{\sigma,n}$ is

$$\|f\|=\sum_{k=1}^{n}\sup_{t\in\mathbb{R}}|e^{\sigma|t|}f^{(k)}(t)|+\sup_{t\le0}|e^{-\sigma t}f(t)|$$
$$+\sup_{t\ge0}|e^{\sigma t}(f(t)-f(\infty))|+|f(\infty)|.$$

THEOREM 2.3. *Let Y denote $(y,\gamma_1,\gamma_2,\cdots,\gamma_N,\delta_1,\delta_2,\cdots,\delta_N)$ and*

$$Y_0=(1,K_1/\lambda_1,\cdots,K_N/\lambda_N,K_1/\lambda_1,\cdots,K_N/\lambda_N).$$

Let $\sigma>0$ be small enough (smaller than all the absolute values of the nonzero eigenvalues of A). Then the following holds: If $f\in Y^{\sigma,n}$ has sufficiently small norm, then (2.2) has a solution Y satisfying $Y-Y_0\in X^{\sigma,n+1}\times(Y^{\sigma,n+1})^{2N}$. Y depends smoothly on f.

Proof. When we put $Y-Y_0=Z$, equation (2.2) can be written in the form $G(Z,f)=0$, and G is a smooth mapping from $(X^{\sigma,n+1}\times(Y^{\sigma,n+1})^{2N})\times Y^{\sigma,n}$ into $(Y^{\sigma,n})^{2N+1}$. Moreover, the linearization $D_ZG(0,0)$ is the mapping $(y,\gamma_i,\delta_i)\to(-\mu\dot{y}+\Sigma_{i=1}^N(\gamma_i-\delta_i)$, $\dot{\gamma}_i+\lambda_i\gamma_i+(2K_i/\mu\lambda_i)\Sigma_j(\gamma_j-\delta_j)$, $\dot{\delta}_i+\lambda_i\delta_i-(K_i/\mu\lambda_i)\Sigma_j(\gamma_j-\delta_j))$. If the γ- and δ-components alone are considered, then, according to Lemma 2.1, all the eigenvalues of the linearization are negative. It is easy to see that, if σ is chosen less than the absolute values of all these eigenvalues, then this part of the operator $D_zG(0,0)$ constitutes an isomorphism from $(Y^{\sigma,n+1})^{2N}$ onto $(Y^{\sigma,n})^{2N}$ (this follows from the fact that $(d/dt+\beta)$ is an isomorphism from $Y^{\sigma,n+1}$ onto $Y^{\sigma,n}$, if β lies outside the interval $(-\sigma,\sigma)$). Moreover, the mapping $y\to\dot{y}$ is a bijection from $X^{\sigma,n+1}$ on $Y^{\sigma,n}$. Therefore $D_zG(0,0)$ is an isomorphism from $X^{\sigma,n+1}\times(Y^{\sigma,n})^{2N}$ onto $(Y^{\sigma,n})^{2N+1}$. The implicit function theorem yields the result. □

Global behavior of solutions for large f.

THEOREM 2.4. *Let $\mu>0$ and $f:\mathbb{R}\to\mathbb{R}$ be continuous and such that $\lim_{t\to-\infty}e^{-\sigma t}f(t)=0$ ($\sigma>0$ is as in Theorem 2.3), $f(t)=0$ for $t\geq t_0$. For every such f, (2.2) has a unique solution satisfying $\lim_{t\to-\infty}y(t)=1$, $\lim_{t\to-\infty}\gamma_i=\lim_{t\to-\infty}\delta_i=K_i/\lambda_i$. This solution exists globally for all times t, and $\lim_{t\to+\infty}y(t)$ exists and is strictly positive.*

Proof. If t_1 is chosen large enough, $e^{-\sigma t}f(t)$ becomes small on $(-\infty,-t_1]$, and one can use an implicit function argument analogous to Theorem 2.3 to prove the existence of a solution on $(-\infty,-t_1]$. This solution is unique in the class of solutions approaching their limiting values at $t=-\infty$ at a rate of $e^{\sigma t}$. However, if a solution tends to these limits at all, it can be seen from the last two equations of (2.2) and the implicit function theorem that γ_i and δ_i tend to their limiting values at a rate of $e^{\sigma t}$. The first equation then implies that y approaches its limiting value at the same rate. Hence the solution is actually unique in the class of all solutions approaching the prescribed limits as $t=-\infty$, as claimed in the theorem.

We now continue this solution to the right, and we have to make sure that it does not blow up at a finite time. For that purpose it is more convenient to consider (2.1) rather than (2.2). From the second and third equation we see that as long as y stays positive, g_i and h_i have a positive lower bound for all finite times, which is independent of $y(t)$. Hence, if y becomes too large, g_iy^3 will dominate over fy^α and also over h_i (since this is less than some constant times $\max_{(-\infty,t]}y(\tau)$). Analogously, if y becomes too small, h_i will dominate over fy^α and g_iy^3. Hence y cannot go to zero or infinity in finite time, whence we find global existence.

Let now $t>t_0$. Then $f=0$, and using (2.2) again, we find

$$\sum_{i=1}^{N}\left[\frac{\mu}{2}\frac{\alpha_i\dot{\alpha}_i}{\alpha_i+\frac{K_i}{\lambda_i}}+\mu\frac{\beta_i\dot{\beta}_i}{\beta_i+\frac{K_i}{\lambda_i}}\right]$$

$$=-\sum_{i=1}^{N}\left[\frac{\lambda_i\mu}{2}\frac{\alpha_i^2}{\alpha_i+\frac{K_i}{\lambda_i}}+\lambda_i\mu\frac{\beta_i^2}{\beta_i+\frac{K_i}{\lambda_i}}\right]-\left[\sum_{i=1}^{N}(\alpha_i-\beta_i)\right]^2. \tag{2.4}$$

Here we have put $\alpha_i=\gamma_i-K_i/\lambda_i$, $\beta_i=\delta_i-K_i/\lambda_i$. As we know that γ_i and δ_i stay positive, the denominators α_i+K_i/λ_i, β_i+K_i/λ_i are always positive. The left-hand side of the equation is the time derivative of a function $F(\alpha_i,\beta_i)$, which, in the range $\alpha_i,\beta_i>-K_i/\lambda_i$, is bounded from below and has a unique nondegenerate minimum at

$\alpha_i=\beta_i=0$. Since F is only determined up to an additive constant, we may normalize it so that F is zero at the minimum. On the other hand, the right-hand side is strictly negative, as long as $(\alpha_1,\cdots,\alpha_N, \beta_1,\cdots,\beta_N)$ stays outside any given neighborhood of 0. As an immediate consequence, we find $\alpha_i\to 0$, $\beta_i\to 0$ as $t\to\infty$ (see e.g. [13, p. 109]). Equation (2.4) has the form $dF/dt=-G$, and in a neighborhood of 0 we have an estimate of the form $G\geq cF$. Hence the convergence of F, and therefore of α_i, β_i to 0 is exponential. One easily concludes from (2.2) that $\ln y$ approaches a constant, and hence $\lim_{t\to\infty} y(t)>0$ exists. □

Remark. A comparison of these results with those of Petrie [17], [18] is interesting. Petrie considers the equation

$$\int_{-\infty}^{t} e^{-\lambda(t-s)}\left(\frac{y^{3-2\nu}(t)}{y^{2-2\nu}(s)}-y^{1-\nu}(s)y^{\nu}(t)\right)ds=f(t)y^2(t).$$

It can be shown that certain generalizations of the rubberlike liquid theory lead to such an equation [14], [15], [16], [19], [20]. Our global existence argument fails if $\nu>\frac{1}{2}$, and, in fact, Petrie has shown that, in this case, solutions can blow up in finite time. It also seems remarkable in this context that, if inertia is included, there is another difference between the cases $\nu<\frac{1}{2}$ and $\nu>\frac{1}{2}$. Whereas $\nu<\frac{1}{2}$ leads to a hyperbolic equation (for $\mu=0$), the type of the equation for $\nu>\frac{1}{2}$ may change to elliptic, thus suggesting the possibility of a very strong spatial instability.

The next corollary provides information on the final recovery for physically significant forces.

COROLLARY 2.5. *Let all the assumptions of Theorem* 2.4 *hold and let y be the solution considered there. If f is always nonnegative and not identically zero, then* $y(\infty)>y(-\infty)=1$; *if f is always nonpositive and not identically zero, then* $y(\infty)<y(-\infty)=1$.

Proof. Assume $f\geq 0$; the other case is analogous. It is immediate from the integral equation (1.1) that $f\geq 0$ implies $y\geq 1$ for all t. Moreover, if $f\neq 0$, there must be some t^* such that $y(t^*)>1$. Let now $z(t)=\min_{\tau\in[t^*,t]} y(\tau)$. Then (1.1) implies that

$$\left[\frac{d}{dt}\right]_+ z(t)\geq\min\left(0,-\frac{1}{\mu}\int_{-\infty}^{t^*}a(t-s)\left(z^3(t)-1\right)ds\right)$$
$$\geq-\frac{1}{\mu}\int_{-\infty}^{t^*}a(t-s)\left(z^3(t)-1\right)ds.$$

If $z(t)-1$ is sufficiently small, this gives us an inequality of the form

$$\left[\frac{d}{dt}\right]_+ z(t)\geq -Ce^{-kt}(z-1).$$

It follows immediately that $\lim_{t\to+\infty} z(t)>1$. □

Remark. With "$>$" or "$<$" replaced by "$\geq$" or "$\leq$", these results are obviously expected on a physical basis. Namely, they simply state that pulling the filament effectively increases its length ($f\geq 0$, $\alpha=2$) or the thickness of the sheet decreases ($f\leq 0$, $\alpha=\frac{1}{2}$), respectively. We have shown that the equal sign never holds, i.e., that the filament can never recover its original length, nor the sheet its original shape.

We now give an argument showing that Theorem 2.4 does not hold if the condition $f(t)=0$ for $t>t_0$ is replaced by exponential decay of f and $\alpha\neq 1$ (in case $\alpha=1$ the previous argument still goes through, the only difference being that $f(t)\ \Sigma_{i=1}^{N}(\alpha_i-\beta_i)$

has to be added on the right-hand side of (2.4)). We restrict ourselves to the case $N=1$. (2.1) reads:

$$-\mu\dot{y}=gy^3-h-f(t)y^\alpha,\quad \dot{g}=-\lambda g+\frac{K}{y^2},\quad \dot{h}=-\lambda h+Ky.$$

We solve these equations for $t\geq 0$ by the following ansatz:

$$y=y_0e^{\nu t},\quad g=g_0e^{-2\nu t}+g_1e^{-\lambda t},\quad h=h_0e^{\nu t}+h_1e^{-\lambda t},$$
$$f=f_0e^{(1-\alpha)\nu t}+g_1y_0^{3-\alpha}e^{((3-\alpha)\nu-\lambda)t}-h_1y_0^{-\alpha}e^{(-\alpha\nu-\lambda)t}.$$

After some calculation one finds that this ansatz satisfies the equations if

$$g_0=\frac{K}{y_0^2(\lambda-2\nu)},\quad h_0=\frac{Ky_0}{\lambda+\nu}\quad\text{and}\quad f_0y_0^{\alpha-1}=\frac{3\nu K+\mu\nu(\lambda-2\nu)(\lambda+\nu)}{(\lambda-2\nu)(\lambda+\nu)}.$$

We thus find solutions for which f goes to zero exponentially, but $y\to\infty$ for $\alpha>1$ and $y\to 0$ for $\alpha<1$.

We have to make sure that, by appropriate continuation for $t<0$, we can match the conditions at $t=-\infty$. For this purpose, we continue y in an arbitrary way to the left such that y is smooth and approaches 1 exponentially at $t=-\infty$. The equations for g and h then have unique solutions approaching K/λ for $t\to-\infty$. These solutions can be matched to the solutions for $t>0$ by appropriate choices of g_1 and h_1. Finally f is determined by the first equation.

The case $\mu=0$. In this case the first equation of (2.1) becomes

$$y^3\sum_{i=1}^N g_i-\sum_{i=1}^N h_i-f(t)y^\alpha=0.$$

PROPOSITION 2.6. *For any $g>0$, $h>0$ and $0<\alpha<3$, the equation $F(y)=gy^3-h-fy^\alpha=0$ has a unique solution in $(0,\infty)$.*

Proof. We have $F(0)<0$, $\lim_{y\to\infty}F(y)>0$, so there is clearly a positive solution. To show it is unique, we investigate zeros of $F'(y)$. We have $F'(y)=3gy^2-\alpha fy^{\alpha-1}$. If $y>0$ and $F'(y)=0$, we find $F(y)=(1/\alpha)yF'(y)+y^3(1-3/\alpha)g-h<0$. This means F cannot have a positive maximum, whence the result. □

The solution $y(g,h,f)$ can then be inserted into the other equations, yielding a system of $2N$ equations.

THEOREM 2.7. *The same statement as in Theorem* 2.4 *holds also for $\mu=0$. Also, Corollary* 2.5 *still holds.*

Sketch of the proof. The existence of a solution on $(-\infty,-t_1]$ and global existence in time are proved in the same manner as before, and we do not repeat the arguments. If $f=0$, one finds from (2.2)

$$\dot{\gamma}_i=-\lambda_i\gamma_i+K_i+2\gamma_i\frac{\dot{y}}{y},\qquad \dot{\delta}_i=-\lambda_i\delta_i+K_i-\delta_i\frac{\dot{y}}{y}.$$

This leads to

$$\sum_{i=1}^N\left[\frac{1}{2}\frac{\alpha_i\dot{\alpha}_i}{\alpha_i+\frac{K_i}{\lambda_i}}+\frac{\dot{\beta}_i\beta_i}{\beta_i+\frac{K_i}{\lambda_i}}\right]=-\sum_{i=1}^N\left[\frac{\lambda_i\alpha_i^2}{2\left(\alpha_i+\frac{K_i}{\lambda_i}\right)}+\frac{\lambda_i\beta_i^2}{\beta_i+\frac{K_i}{\lambda_i}}\right]+\frac{\dot{y}}{y}\sum_{i=1}^N(\alpha_i-\beta_i),$$

where α_i and β_i are defined as before.

Since $\Sigma_i(\alpha_i-\beta_i)$ is now equal to zero, we still find that α_i and β_i approach 0 exponentially, whence the result.

For the corollary, observe that

$$\dot{y}(t)\cdot 3y^2(t)\int_{-\infty}^{t} a(t-s)\frac{1}{y^2(s)}=-\int_{-\infty}^{t} a'(t-s)\left[\frac{y^3(t)}{y^2(s)}-y(s)\right]ds.$$

Using this, one can apply an argument analogous to the previous one.

Finally, we want to prove that solutions depend continuously on μ, even at $\mu=0$. Let $f\in Y^{\sigma,n}$ be given such that it either has a small norm or it satisfies the conditions of Theorem 2.4. We know that a unique solution $y(t)$ satisfying $y(-\infty)=1$ exists both for $\mu=0$ and for $\mu>0$. In (2.1), we put $g=\Sigma_{i=1}^{N} g_i$, $h=\Sigma_{i=1}^{N} h_i$, and $z=y-\sqrt[3]{h/g}$ (for $\mu=0$, $f=0$, the first equation of (2.1) is solved by $y=\sqrt[3]{h/g}$). We obtain

$$-\mu\dot{z}=g\left(\sqrt[3]{\frac{h}{g}}+z\right)^3-h-f(t)\left(\sqrt[3]{\frac{h}{g}}+z\right)^{\alpha}+\mu\frac{d}{dt}\sqrt[3]{\frac{h}{g}},$$

(2.5)

$$\dot{g}_i=-\lambda_i g_i+\frac{K_i}{\left(\sqrt[3]{\frac{h}{g}}+z\right)^2},\qquad \dot{h}_i=-\lambda_i h_i+K_i\left(\sqrt[3]{\frac{h}{g}}+z\right).$$

As we have proved, there exists some $\mu_0>0$ such that for every $\mu\in[0,\mu_0]$, system (2.5) has a unique solution in the Banach manifold

$$M_n=\left\{(z,g_i,h_i)\,\middle|\, z\in Y^{\sigma,n},\ g_i-\frac{K_i}{\lambda_i\sqrt[3]{\frac{h^2}{g^2}}}\in Y^{\sigma,n},\ h_i-\frac{K_i}{\lambda_i}\sqrt[3]{\frac{h}{g}}\in Y^{\sigma,n}\right\}.$$

In particular, let z_0, $g_{i,0}$, $h_{i,0}$ denote the solution for $\mu=0$.

Linearizing at this solution (or likewise at any solution for $\mu>0$), we obtain a system of linear ODEs with a matrix approaching a constant limit as $t\to-\infty$ and $t\to+\infty$. From a discussion of the asymptotic behavior of solutions of the linearized system for $t\to\pm\infty$, one can easily see that for any inhomogeneity in $(Y^{\sigma,n})^{2N+1}$ there is a unique solution in the tangent space of M_n. The argument parallels our existence proof for solutions: First consider the problem on $(-\infty,-T]$ with T large, where the matrix is approximated by the linearization at the trivial solution. Continuation of solutions for $t>-T$ presents no problem, since the equation is linear, and finally the behavior for $t\to+\infty$ must be discussed. We leave the details of the analysis to the reader.

Thus the linearization is a densely defined bijective operator from the tangent space of M_n into $(Y^{\sigma,n})^{2N+1}$. It is thus natural to attempt proving the existence of a continuous family of solutions in a neighborhood of $\mu=0$ using the implicit function theorem. One does, however, face the problem that the term $\mu\dot{z}$ represents an unbounded operator.

The first equation of (2.4) has the form

$$-\mu\frac{d}{dt}(z-z_0)=\rho(t)(z-z_0)-f(t)\cdot L(h-h_0,f-f_0)+\text{nonlinear terms}+O(\mu)$$

where

$$\rho(t)=3g_0\left(\sqrt[3]{\frac{h_0}{g_0}}+z_0\right)^2-\alpha f(t)\left(\sqrt[3]{\frac{h_0}{g_0}}+z_0\right)^{\alpha-1}$$

is positive and converges to $\Sigma_{i=1}^{N} K_i/\lambda_i$ for $t\to\pm\infty$. L is linear in its arguments, and the term $O(\mu)$ does not involve any unbounded operators after the second and third equation of (2.4) have been substituted into the first to replace $\dot{g}$ and $\dot{h}$. It is easy to show that the operator $(\mu d/dt+\rho(t))^{-1}$: $Y^{\sigma,n}\to Y^{\sigma,n}$ is strongly continuous with respect to μ. Denoting $V=(z-z_0, g_1-g_{1,0},\cdots,g_N-g_{N,0}, h_1-h_{1,0},\cdots,h_N-h_{N,0})$, we can thus rewrite (2.5) in the abstract form,

$$L(\mu)V=N(\mu,V)\Leftrightarrow V-(L(\mu))^{-1}N(\mu,V)=0, \tag{2.6}$$

where $L(\mu)$ has a strongly continuous inverse and $N(0,0)=0$, $D_V N(0,0)=0$.

The existence of a continuous solution $V(\mu)$ now follows from the following theorem. □

THEOREM 2.8. *Let X, Y and Z be Banach spaces, U a neighborhood of $(0,0)$ in $X\times Y$, and F: $U\to Z$ a mapping having the following properties*:

(i) $F(0,0)=0$,

(ii) *F is continuous*,

(iii) *F is continuously differentiable with respect to y for each fixed x*,

(iv) *$D_y F(0,0)$: $Y\to Z$ is an isomorphism*,

(v) *$D_y F$ is continuous at the point $(0,0)$*.

Then the equation $F(x,y)=0$ has a unique resolution $y=\varphi(x)$ in some neighborhood of $(0,0)$, and φ is continuous.

The proof of this theorem differs by no means from the standard proof of the implicit function theorem (cf. [10], [11]), but it is crucial for our problem that (iii) and (v) are sufficient, rather than continuity of $D_y F$ in a neighborhood of $(0,0)$, as usually required. Namely, we can identify X with $\mathbf{R}$, Y with the tangent space of M_n, Z with $Y^{\sigma,n}$, with μ and y with V. For μ fixed, the term $L(\mu)^{-1}N(\mu,V)$ depends smoothly on V; moreover, since $\lim_{\mu\to 0,V\to 0} D_V N(\mu,V)=0$, we also have

$$\lim_{\mu\to 0,V\to 0} D_V\big(L(\mu)^{-1}N(\mu,V)\big)=\lim_{\mu\to 0,V\to 0} L(\mu)^{-1}D_V N(\mu,V)=0.$$

Hence Theorem 2.8 applies to (2.6), although the standard form of the implicit function theorem would not. This yields a continuous solution $V=V(\mu)$.

Moreover, the mapping $(\mu,z)\mapsto(\mu d/dt+\rho(t))^{-1}z$ is a C^k-mapping from $\mathbf{R}\times Y^{\sigma,n}$ into $Y^{\sigma,n-k}$. From the following theorem, which was also proved in [10], [11], one concludes that $V(\mu)$ is actually a C^k-function of μ when regarded as lying in $Y^{\sigma,n-k}$.

THEOREM 2.9. *Let $Y^{(k)}$ and $Z^{(k)}$ respectively $(k=0,1,\cdots,N)$ be two hierarchies of Banach spaces such that $Y^{(k)}\subset Y^{(k+1)}$, $Z^{(k)}\subset Z^{(k+1)}$, the imbeddings being continuous. Let X be a finite dimensional Banach space and F a mapping from a neighborhood U of 0 in $X\times Y^{(N)}$ into $Z^{(N)}$ having the following properties*:

(i) $F(U\cap(X\times Y^{(k)}))\subset Z^{(k)}$, $k=0,1,\cdots,N$.

(ii) *For each fixed k, $F_k := F|_{U\cap(X\times Y^{(k)})}$ satisfies the conditions of Theorem 2.8, when it is considered as a mapping from $X\times Y^{(k)}$ into $Z^{(k)}$. For x fixed, $F_k(x,\cdot)$ is a smooth (i.e. sufficiently often differentiable) mapping.*

(iii) *F: $X\times Y^{(k)}\to Z^{(k+m)}$ is of class C^m for each $k=0,1,\cdots,N$ and $m\le N-k$.*

(iv) *The mapping $(x,y,u^1,\cdots,u^j)\to z=D_{x^i y^j}F(x,y)(u^1,\cdots,u^j)$ is continuous from $X\times Y^{(k)}\times(Y^{(k)})^j$ into $\mathfrak{L}^i(X,Z^{(k+i)})$.*

Then the solution $y=\varphi(x)\in Y^{(0)}$ existing by Theorem 2.8 is a C^m-function of x in some neighborhood V_m of 0, if y is regarded as an element of $Y^{(m)}$.

We summarize our results in the following:

THEOREM 2.10. *Let $f \in Y^{\sigma,n}$ be given such that either f has small norm or $f(t) \equiv 0$ for t greater than some $t_0 < \infty$. Then, for each $\mu \in [0, \infty]$, (1.1) has a unique solution y satisfying $y - 1 \in X^{\sigma,n}$. In the limit $\mu \to 0$, $y - 1 \in X^{\sigma,n}$ depends continuously on μ, and it is a C^k-function of μ when regarded as dwelling in $X^{\sigma,n-k}$.*

3. The discretization scheme. When solving (1.1) numerically, one faces the problem that it is to be solved on an infinite interval. A reasonable way of doing this is to cut at $-T \ll 0$, and replace $y(t)$ for $t \leq -T$ by its limit $\lim_{t \to -\infty} y(t) = 1$. We thus obtain the approximating problem

$$\mu \dot{y}_{-T} + \int_{-\infty}^{-T} a(t-s)\,ds \cdot \left(y_{-T}^3(t) - 1\right) + \int_{-T}^{t} a(t-s)\left(\frac{y_{-T}^3(t)}{y_{-T}^2(s)} - y_{-T}(s)\right) ds - f(t) y_{-T}^{\alpha}(t) = 0, \tag{3.1}$$

$$y_{-T}(t) - 1 = 0, \qquad t \leq -T. \tag{3.2}$$

On the finite interval the integrodifferential equation can now be discretized in a straightforward manner. Like Nevanlinna [8], we use a first order implicit (Euler-type) method, because for this simple scheme we can prove that the qualitative properties of solutions of (1.1), such as exponential decay at infinity and uniform convergence as $\mu \to 0$, carry over to the discrete problem. Since these properties are essential for the continuous problems it is very reasonable to require that the computed approximating solutions exhibit them too. Our computations have shown that good approximations can be obtained with quite large mesh sizes, and so the computational effort for the first order scheme remains reasonably small.

We choose mesh points $t_i = ih$, $i \in \mathbf{Z}$, where $t_{-m} = -T$, and denote by y_i the approximation to $y(t_i)$ (or to $y_{-T}(t_i)$). Then the discretized form of the equation reads

$$\mu \frac{y_i - y_{i-1}}{h} + \int_{-\infty}^{t_{-m}} a(t_i - s)\,ds \cdot \left(y_i^3 - 1\right) + h \sum_{j=-m+1}^{i} a((i-j)h)\left(\frac{y_i^3}{y_j^2} - y_j\right) - f(t_i) y_i^{\alpha} = 0, \qquad i > -m, \tag{3.3}$$

$$y_i - 1 = 0, \qquad i \leq -m. \tag{3.4}$$

Obviously,

$$\int_{-\infty}^{t_{-m}} a(t_i - s)\,ds = \sum_{l=1}^{N} \frac{K_l}{\lambda_l} e^{\lambda_l (t_{-m} - t_i)}. \tag{3.5}$$

Equation (3.3) has the form

$$c_1 y_i^3 + c_2 y_i^{\alpha} + c_3 y_i = c_4, \tag{3.6}$$

where the c's depend on μ, h, t_{-m}, t_i and $y_j, j < i$.[1]

[1] Equation (3.6) will be discussed at the end of this section.

The analysis of the discrete equation will be carried out in the same sort of spaces as the analysis of the continuous equation. We therefore define discrete analogues of the exponentially weighted spaces introduced in Definition 2.2.

DEFINITION 3.1. Let

$$X_h^\sigma := \left\{ \hat{f}=(f_i)_{i=-\infty}^{\infty} \in l^\infty \,\middle|\, \lim_{i\to\infty} f_i =: f_\infty \text{ exists}, \lim_{i\to\infty} e^{i\sigma h}|f_i - f_\infty| = 0, \lim_{i\to-\infty} e^{-i\sigma h}|f_i| = 0 \right\}$$

and

$$Y_h^\sigma := \left\{ \hat{g}=(g_i)_{i=-\infty}^{\infty} \in l^\infty \,\middle|\, \lim_{i\to\infty} e^{i\sigma h}|g_i| = \lim_{i\to-\infty} e^{-i\sigma h}|g_i| = 0 \right\}.$$

The natural norms in these spaces are

$$\|\hat{f}\|_{X_h^\sigma} = \sup_{i>0} e^{i\sigma h}|f_i - f_\infty| + \sup_{i\le 0} e^{-i\sigma h}|f_i| + |f_\infty|$$

and

$$\|\hat{g}\|_{Y_h^\sigma} = \sup_{i>0} e^{i\sigma h}|g_i| + \sup_{i\le 0} e^{-i\sigma h}|g_i|.$$

Setting $\hat{z}=(y_i-1)_{i=-\infty}^{\infty}$, we rewrite (3.3) and (3.4) in the abstract form

$$F_{h,m}(\hat{z}) = 0. \tag{3.7}$$

Now let $f \in Y^{\sigma,n}$ ($\sigma>0$, $n\in N$) be given such that the assumptions of Theorem 2.4 hold. It is an easy exercise to show that, for any $\varepsilon \in [0,\sigma)$,

$$F_{h,m}: X_h^{\sigma-\varepsilon} \to Y_h^{\sigma-\varepsilon} \tag{3.8}$$

(we explain below why ε is introduced).

The aim of the following analysis is to prove that $(y_i)_{i=-\infty}^{\infty}$ converges to $(y(t_i))_{i=-\infty}^{\infty}$ in the topology of $X_h^{\sigma-\varepsilon}$. The proof will be based on Keller's [3] nonlinear stability-consistency concept.

Let us first show consistency. The local discretization error $l=(l_i)_{i=-\infty}^{\infty}$ is defined by

$$l = F_{h,m}\big((y(t_i)-1)_{i=-\infty}^{\infty}\big). \tag{3.9}$$

For $f\in Y^{\sigma,1}$ (which implies $\mu y \in Y^{\sigma,2}$, uniformly in μ), we find, using the exponential decay of y as $t\to\pm\infty$, that in the limit $t_{-m}\to-\infty$, $h\to 0$

$$\text{(a)} \qquad |l_i| \le o(1)e^{-\sigma|t_i|}, \qquad i\le -m,$$

$$\text{(b)} \qquad \left|\mu\frac{y(t_i)-y(t_{i-1})}{h} - \mu y'(t_i)\right| = \text{const.}\, o(1) h e^{-\sigma|t_i|},$$

$$\text{(c)} \qquad \left|\int_{-\infty}^{t_{-m}} a(t_i-s)\,ds\cdot\big(y^3(t_i)-1\big) - \int_{-\infty}^{t_{-m}} a(t_i-s)\left(\frac{y^3(t_i)}{y^2(s)} - y(s)\right) ds\right| \le \text{const.}\, o(1) e^{\sigma(2t_{-m}-t_i)}, \tag{3.10}$$

$$\text{(d)} \qquad \left|h\sum_{j=-m+1}^{i} a((i-j)h)\left(\frac{y^3(t_i)}{y^2(t_j)} - y(t_j)\right) - \int_{t_{-m}}^{t_i} a(t_i-s)\left(\frac{y^3(t_i)}{y^2(s)} - y(s)\right) ds\right| \le \text{const.}\, o(1) h e^{-\sigma|t_i|}.$$

Here $o(1)$ stands for a factor that vanishes as $t_i \to -\infty$. Therefore,

$$|l_i| \le \text{const.}\, o(1) \begin{cases} h e^{-\sigma|t_i|} + e^{-\sigma|t_i - 2t_{-m}|}, & i > -m, \\ e^{-\sigma|t_i|}, & i \le -m. \end{cases} \tag{3.11}$$

From Definition 3.1 we conclude that

$$\|l\|_{Y_h^{\sigma-\varepsilon}} \le \text{const.}\big(h + o(e^{-\varepsilon|t_{-m}|})\big). \tag{3.12}$$

The constant is independent of h, t_{-m}, $0 \le \mu < \infty$, $0 \le \varepsilon \le \varepsilon_0 < \sigma$. Note that, in particular, the estimate contains a term $o(e^{-\varepsilon|t_{-m}|})$. The reason for this is that, when approximating (1.1) by (3.1), (3.2), we have replaced f by 0 for $t \le -T$, and in the norm of $Y^{\sigma-\varepsilon,n}$ this introduces an error of the order $o(e^{-\varepsilon|T|})$. This is the reason why we have introduced ε; for $\varepsilon = 0$ we would still get convergence, but no estimate for the order. (3.12) settles consistency.

For the stability analysis, we calculate the Fréchet derivative of $F_{h,m}$ at the exact solution $(y(t_i) - 1)_{i=-\infty}^{\infty}$, which is denoted by

$$L_{h,m} := D_{\hat{z}} F_{h,m}\big((y(t_i) - 1)_{i=-\infty}^{\infty}\big). \tag{3.13}$$

For $\hat{u} = (u_i)_{i=-\infty}^{\infty} \in X_h^{\sigma-\varepsilon}$ we obtain

$$(L_{h,m}\hat{u})_i = \begin{cases} u_i, & i \le -m, \\ \mu \dfrac{u_i - u_{i-1}}{h} + 3\displaystyle\int_{-\infty}^{t_{-m}} a(t_i - s)\, ds\, y(t_i)^2 u_i \\ \quad + 3h\left(\displaystyle\sum_{j=-m+1}^{i} a((i-j)h) \frac{y^2(t_i)}{y^2(t_j)} \right) u_i \\ \quad - h \displaystyle\sum_{j=-m+1}^{i} a((i-j)h) \left(2\frac{y^3(t_i)}{y^3(t_j)} + 1 \right) u_j \\ \quad - \alpha f(t_i) y(t_i)^{\alpha-1} u_i, & i > -m. \end{cases} \tag{3.14}$$

Stability means that $L_{h,m}^{-1}$ exists and that it is bounded as an operator from $Y_h^{\sigma-\varepsilon}$ into $X_h^{\sigma-\varepsilon}$ uniformly with respect to h, t_{-m}, μ and $0 \le \varepsilon \le \varepsilon_0 < \sigma$. Therefore we look at the equation $L_{h,m}\hat{u} = \hat{v}$ for $\hat{v} = (v_i)_{i=-\infty}^{\infty} \in Y_h^{\sigma-\varepsilon}$.

For $i \le -m$ we find $u_i = v_i$, and for $i > -m$ we show that

$$\begin{aligned} G_i(h, -m, \mu) = \frac{\mu}{h} + 3\int_{-\infty}^{t_{-m}} a(t_i - s) y^2(t_i)\, ds + 3h \sum_{j=-m+1}^{i-1} a((i-j)h) \frac{y^2(t_i)}{y^2(t_j)} \\ - \alpha f(t_i) y^{\alpha-1}(t_i), \end{aligned} \tag{3.15}$$

which is the coefficient of u_i in (3.14), is bounded away from 0 uniformly in h, t_{-m} and μ.

It is easy to show that

$$\begin{aligned} G_i(h, -m, \mu) = \frac{\mu}{h} + 3\int_{-\infty}^{t_i} a(t_i - s) \frac{y^2(t_i)}{y^2(s)}\, ds - \alpha f(t_i) y^{\alpha-1}(t_i) \\ + O(h) + O(e^{-\sigma(|t_i| + |t_{-m}|)}). \end{aligned} \tag{3.16}$$

Using (1.1) we get

(3.17)
$$G_i(h,-m,\mu)=\frac{\mu}{h}+\frac{3}{y(t_i)}\left(\int_{-\infty}^{t_i}a(t_i-s)y(s)\,ds-\mu\dot{y}(t_i)\right.$$
$$\left.+\left(1-\frac{\alpha}{3}\right)f(t)y^{\alpha}(t_i)\right)+O(h)+O(e^{-\sigma(|t_i|+|t_{-m}|)}).$$

It follows from §2 that

(3.18)
$$0<\underline{y_0}\le y(t,\mu)\le\bar{y_0},\qquad |\dot{y}(t,\mu)|\le\tilde{y}$$

uniformly for $\mu\in[0,\infty)$. For $f\le 0$, (3.16) provides a uniform lower bound for G_i, and, for $f\ge 0$, (3.18) provides a uniform lower bound, since $\alpha<3$.

The preceding considerations make it apparent, why the term $\int_{-\infty}^{t_{-m}}a(t_i-s)\,ds(y_i^3-1)$ should be maintained in (3.3). If this term were neglected, the uniform bounds on G_i would no longer hold, and, unless a constraint of the form $\mu/h\ge$const. is imposed, an artificial boundary layer can be generated at t_{-m}.

We see from the above that $L_{h,m}$ can formally be inverted. It remains to be proved that the solution $\hat{u}$ of

(3.19)
$$L_{h,m}\hat{u}=\hat{v}$$

satisfies an estimate

(3.20)
$$\|\hat{u}\|_{X_h^{q-\varepsilon}}\le\text{const.}\|\hat{v}\|_{Y_h^{q-\varepsilon}}$$

with the constant independent of h,t_{-m},μ and ε.

We begin with the reduced problem for $\mu=0$. Equation (3.14) yields

(3.21)
$$u_i=h\sum_{j=-m+1}^{i-1}a((i-j)h)\alpha_{i,j}(h,-m)u_j+v_i^0,$$

where

(3.22)
$$\alpha_{i,j}(h,-m)=\frac{2y^3(t_i)/y^3(t_j)+1}{G_i(h,-m,0)},\qquad v_i^0=\frac{v_i}{G_i(h,-m,0)},$$

with G_i as defined in (3.17). Since $y(-\infty)=1$, this implies, after a simple calculation,

(3.23)
$$\alpha_{i,j}(h,-m)=\frac{1}{h\sum_{l=1}^{N}K_l/(e^{\lambda_l h}-1)}+\gamma=:C(h,\gamma)$$

where $\gamma\to 0$ as $h\to 0$, $t_{-m}\to-\infty$, $t_i\to-\infty$.

Using the form (1.2) of a, we get from (3.21)

(3.24)
$$|u_i|\le C(h,\gamma)h\sum_{j=-m+1}^{i-1}\sum_{l=1}^{N}K_l e^{-\lambda_l(t_i-t_j)}|u_j|+|v_i^0|.$$

The solution w_i of the equation obtained by replacing $|v_i|$ by the larger quantity $e^{i(\sigma-\varepsilon)h}\|\bar{v}\|_{l^\infty}$, where

(3.25)
$$v_i^0=e^{i(\sigma-\varepsilon)h}\bar{v}_i,\qquad \bar{v}=(\bar{v}_i)_{i=-\infty}^{\infty}\in l^\infty,$$

provides an upper bound for $|u_i|$. In analogy to §2, we substitute

(3.26)
$$g_i^l=K_l h\sum_{j=-m+1}^{i-1}e^{-\lambda_l(t_i-t_j)}w_j,\qquad l=1,2,\cdots,N,$$

which yields the difference equation

$$g_{i+1}^l - g_i^l = K_l h e^{-\lambda_l h} w_i + (e^{-\lambda_l h} - 1) g_i^l \tag{3.27}$$

and the following difference equation for w_i:

$$w_{i+1} - w_i = C(h,\gamma) \sum_{l=1}^{N} (e^{-\lambda_l h} - 1) g_i^l + C(h,\gamma) \sum_{l=1}^{N} K_l h e^{-\lambda_l h} w_i + (e^{(\sigma-\varepsilon)h} - 1) e^{i(\sigma-\varepsilon)h} \|\bar{v}\|_{l^\infty}. \tag{3.28}$$

Formulas (3.27) and (3.28) form a system of difference equations. Setting

$$z_i = (w_i, g_i^1, \cdots, g_i^N), \qquad \varphi_i = ((e^{(\sigma-\varepsilon)h} - 1) e^{i(\sigma-\varepsilon)h} \|\bar{v}\|_{l^\infty}, 0, \cdots, 0), \tag{3.29}$$

we can rewrite this system in the form

$$z_{i+1} = (I + A(h,\gamma)) z_i + \varphi_i, \qquad i \geq -m, \tag{3.30}$$

where

(3.31)

$$A(h,\gamma) = \begin{bmatrix} C(h,\gamma) \sum_{l=1}^{N} K_l h e^{-\lambda_l h} & C(h,\gamma)(e^{-\lambda_1 h} - 1) & \cdots & C(h,\gamma)(e^{-\lambda_N h} - 1) \\ K_1 h e^{-\lambda_1 h} & e^{-\lambda_1 h} - 1 & & \\ \vdots & & \ddots & 0 \\ K_N h e^{-\lambda_N h} & 0 & & e^{-\lambda_N h} - 1 \end{bmatrix}$$

The solution of (3.30) is given by

$$z_i = (I + A(h,\gamma))^{i-1+m} z_{-m+1} + \sum_{j=-m}^{i-1} (I + A(h,\gamma))^{i-j-1} \varphi_j, \tag{3.32}$$

with the initial condition $z_{-m+1} = (e^{-(m-1)(\sigma-\varepsilon)h} \|\bar{v}\|_{l^\infty}, 0, \cdots, 0)$.

The goal of the following analysis is to show that

$$\sup_{i \leq 0} e^{-i(\sigma-\varepsilon)h} \|z_i\| \leq \text{const.} \|\bar{v}\|_{l^\infty}, \tag{3.33}$$

which implies

$$\sup_{i \leq 0} e^{-i(\sigma-\varepsilon)h} |u_i| \leq \text{const.} \sup_{i \leq 0} e^{-i(\sigma-\varepsilon)h} |v_i|. \tag{3.34}$$

Summing up the geometric series in (3.32), we obtain the estimate

$$e^{-i(\sigma-\varepsilon)h} \|z_i\| \leq \text{const.} \left[\frac{\left\| (I + A(h,\gamma))^{i+m-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\|}{e^{(i+m-1)(\sigma-\varepsilon)h}} + 1 \right] \|\bar{v}\|_{l^\infty} \cdot \left[((e^{(\sigma-\varepsilon)h} - 1)) \left\| \left(I - \frac{I + A(h,\gamma)}{e^{(\sigma-\varepsilon)h}} \right)^{-1} \right\| \right]. \tag{3.35}$$

We thus have to prove estimates of the following form:

$$\left\| \left(\frac{I+A(h,\gamma)}{e^{(\sigma-\varepsilon)h}} \right)^k \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\| \leq \text{const.}, \qquad k \in N \tag{3.36}$$

and

$$|e^{(\sigma-\varepsilon)h}-1| \cdot \left\| \left(I - \frac{I+A(h,\gamma)}{e^{(\sigma-\varepsilon)h}} \right)^{-1} \right\| \leq \text{const.} \tag{3.37}$$

Both will follow from an analysis of the Jordan form of $A(h,\gamma)$. For h,γ small enough we can write

$$A(h,\gamma) = h\left(\frac{d}{dh} A(0,0) + O(h) + O(\gamma) \right) \tag{3.38}$$

where

$$\frac{d}{dh}A(0,0) = \begin{pmatrix} C(0,0)\sum_{l=1}^{N} K_l & -C(0,0)\lambda_1 & \cdots & -C(0,0)\lambda_N \\ K_1 & -\lambda_1 & & 0 \\ \vdots & 0 & \ddots & \\ K_N & & & -\lambda_N \end{pmatrix}. \tag{3.39}$$

This matrix has the characteristic polynomial

$$p(\rho) = C(0,0)\sum_{l=1}^{N} K_l - \rho - C(0,0)\sum_{l=1}^{N} \frac{\lambda_l K_l}{\lambda_l + \rho}. \tag{3.40}$$

Recalling that $C(0,0) = (\Sigma_{l=1}^{N} K_l/\lambda_l)^{-1}$, it is an easy exercise to show that the root 0 is two-fold. An analysis similar to that given for (2.3) shows that all remaining zeros are real and negative.

A similar calculation shows that zero is also a double eigenvalue of $(d/dh)A(0,0) + O(h)$ (as of (3.38)). Therefore we get for the eigenvalues of $A(h,\gamma)$,

$$\begin{aligned} &\text{(a)} && \rho_1(h,\gamma) = h\,o(1), \quad \rho_2(h,\gamma) = h\,o(1) \quad \text{as } \gamma \to 0, \\ &\text{(b)} && \rho_i(h,\gamma) = h(\tilde{\rho}_i + o(1)) \quad \text{as } h,\gamma \to 0, \quad i = 3(1)(N+1), \end{aligned} \tag{3.41}$$

where $\tilde{\rho}_i < 0$ for $i = 3(1)(N+1)$, such that

$$\left| \frac{1+\rho_i(h,\gamma)}{e^{(\sigma-\varepsilon)h}} \right| \leq 1, \tag{3.42}$$

holds. Equality in (3.42) only holds for $h=0$. Since $(1/h)A(h,0)$ is holomorphic in $h=0$, and since the eigenvalues of $(1/h)A(h,0)$ do not change multiplicities as $h \to 0$ (i.e., the negative eigenvalues of $(d/dh)A(0,0)$ are distinct and 0 is a double eigenvalue of $(1/h)A(h,0)$), there is a matrix $G(h)$ such that $G(h)$, $G^{-1}(h)$ are holomorphic in $h=0$ and $J(h)$ defined by

$$A(h,0) = G(h)J(h)G^{-1}(h)$$

is the Jordan form of $A(h,0)$ (see [2]). Therefore (3.36), (3.37) hold for $\gamma=0$. A simple perturbation argument assures (3.36), (3.37) for γ sufficiently small. Thus for h sufficiently small and $K>0$ sufficiently large, we have proved that

$$\sup_{t_i\le -K} e^{-t_i(\sigma-\varepsilon)}|u_i|\le \text{const.}\|\bar{v}\|_{l^\infty}. \tag{3.43}$$

The solution u_i can be continued over the finite interval $[-K,0]$, and by a standard stability analysis (see [1]) we obtain (3.34).

We now have to treat the case $t_i\ge 0$. For this, we rewrite (3.21) as

$$\begin{aligned} u_i=h&\sum_{j=-m+1}^{I-1} a((i-j)h)\alpha_{i,j}(h,-m)u_j \\ &+h\sum_{j=I}^{i-1} a((i-j)h)\alpha_{i,j}(h,-m)u_j+v_i, \end{aligned} \tag{3.44}$$

where it is assumed that $t_I\ge K$ is sufficiently large. After some calculation, we get from (3.15), (3.22):

$$\alpha_{i,j}(h,-m)=D(h)+\beta_{ij}(h,-m) \tag{3.45}$$

where

$$D(h)=\frac{1}{h\sum_{l=1}^N K_l/(e^{\lambda_l h}-1)}, \qquad |\beta_{i,j}(h,-m)|=O(e^{-\sigma t_j}), \qquad t_j\ge K. \tag{3.46}$$

It is therefore natural to study the equation

$$\tilde{u}_i=D(h)\sum_{j=I}^{i-1} a((i-j)h)\tilde{u}_j+\tilde{v}_i, \tag{3.47}$$

where $\tilde{v}_i$ is v_i plus the first sum in (3.44), and interpret (3.44) as a perturbation of (3.47). As before, we put

$$g_i^l=K_l h\sum_{j=I}^{i-1} e^{-\lambda_l(t_i-t_j)}\tilde{u}_j, \tag{3.48}$$

which leads to the difference equation

$$g_{i+1}^l-g_i^l=K_l D(h)he^{-\lambda_l h}\sum_{j=1}^{N} g_i^j+(e^{-\lambda_l h}-1)g_i^l+hK_l e^{-\lambda_l h}\tilde{v}_i. \tag{3.49}$$

Here the relation

$$\tilde{u}_i=D(h)\sum_{j=1}^{N} g_i^j+\tilde{v}_i \tag{3.50}$$

has been used.

Putting $g_i=(g_i^1,\cdots,g_i^N)$, $e(h):=(K_1e^{-\lambda_1 h},\cdots,K_N e^{-\lambda_N h})$, we obtain the following matrix form of (3.49)

$$g_{i+1}=(I+B(h))g_i+\tilde{v}_i h\cdot e(h). \tag{3.51}$$

This has the solution

$$g_i=h\sum_{j=I}^{i-1}(I+B(h))^{i-j-1}\tilde{v}_j e(h). \tag{3.52}$$

When dealing with the case $t_i<0$, we used a redundant system of equations rather than an analogue of (3.49). The reason for this was that it is easier to compute the characteristic polynomial of the matrix of the "redundant" system. In the "redundant" ($(N+1)$-dimensional instead of N-dimensional) form (3.49) reads

$$\begin{aligned}\tilde{u}_{i+1}-\tilde{u}_i&=D(h)h\sum_{l=1}^{N}K_l e^{-\lambda_l h}\tilde{u}_i+D(h)\sum_{l=1}^{N}(e^{-\lambda_l h}-1)g_i^l+\tilde{v}_{i+1}-\tilde{v}_i,\\ g_{i+1}^l-g_i^l&=K_l h e^{-\lambda_l h}\tilde{u}_i+(e^{-\lambda_l h}-1)g_i^l.\end{aligned}\tag{3.53}$$

When we write this in matrix form

$$\tilde{z}_{i+1}=(I+A(h))\tilde{z}_i+d_i\tag{3.54}$$

(where $z_i=(\tilde{u}_i,g_i^1,\cdots,g_i^N)$, $d_i=(\tilde{v}_{i+1}-\tilde{v}_i,0,\cdots,0)$), we see immediately that $A(h)$ is the same matrix as (3.31), except that $C(h,\gamma)$ is replaced by $D(h)$. The characteristic polynomial is

$$q(\rho)=D(h)h\sum_{l=1}^{N}K_l e^{-\lambda_l h}-\rho+D(h)h\sum_{l=1}^{N}K_l\frac{e^{-\lambda_l h}(e^{-\lambda_l h}-1)}{\rho-(e^{-\lambda_l h}-1)}.\tag{3.55}$$

It is easily verified that $\rho=0$ is a double root. Moreover, since

$$D(h)=C(0,0)+O(h),\tag{3.56}$$

the other roots are small perturbations of those of $p(\rho)$ as given by (3.40), and therefore have negative real parts. When we pass from $A(h)$ to the N-dimensional matrix $B(h)$, the eigenvalues obviously remain the same, except that 0 as an eigenvalue of $B(h)$ has multiplicity one rather than two. Hence there is a matrix $E(h)$ such that $E(h),E^{-1}(h)$ are continuous for $h\in[0,h_0]$ and the Jordan form $J(h)$ of $B(h)$

$$J(h)=E^{-1}(h)B(h)E(h)\tag{3.57}$$

has the block form

$$J(h)=\begin{pmatrix}0 & 0 & \cdots & 0\\ 0 & & & \\ 0 & & & \\ 0 & & J_-(h) & \\ \vdots & & & \\ 0 & & & \end{pmatrix},\tag{3.58}$$

where the $(N-1)\times(N-1)$-matrix $J_-(h)$ has only eigenvalues with negative real parts. The continuity of $E(h)$, $E^{-1}(h)$ holds since $\frac{1}{h}B(h)$ is analytic in h and since the eigenvalues of $\frac{1}{h}B(h)$ are distinct even for $h=0$ (see [2]).

If we put $g_i=E(h)w_i$, (3.52) yields

$$w_i=h\sum_{j=I}^{i-1}(I+J(h))^{i-j-1}\tilde{v}_j E^{-1}(h)e(h).\tag{3.59}$$

In the first component this reads in particular

$$w_i^1=h\sum_{j=I}^{i-1}\tilde{v}_j\big(E^{-1}(h)e(h)\big)^1.\tag{3.60}$$

From this we obtain the following estimate for t_I sufficiently large:

$$\begin{aligned} &\lim_{i\to\infty}|w_i^1|\le \text{const.}\sup_{i\ge I} e^{i(\sigma-\varepsilon)h}|\tilde v_i|,\\ &\sup_{i\ge I} e^{i(\sigma-\varepsilon)h}\Big|w_i^1-\lim_{i\to\infty} w_i^1\Big|\le \text{const.}\, o(1)\sup_{i\ge I} e^{i(\sigma-\varepsilon)h}|\tilde v_i|. \end{aligned} \tag{3.61}$$

For the components $(w_i^2,\cdots,w_i^N)$, where only eigenvalues with negative real parts are involved, an analogous estimate follows from the same arguments that have been used in the case $t_i<0$, and we even have $\lim_{i\to\infty} w_i^l=0$ for $l>1$.

Let us now introduce the spaces

$$A_{h,I}^{\sigma-\varepsilon}=\Big\{f=(f_i)_{i=I}^{\infty}\Big|\lim_{i\ge I} e^{i(\sigma-\varepsilon)h}|f_i|=0\Big\},$$

$$B_{h,I}^{\sigma-\varepsilon}=\Big\{f=(f_i)_{i=I}^{\infty}\Big|\lim_{i\to\infty} f_i=:f_\infty \text{ exists, } \lim_{i\ge I} e^{i(\sigma-\varepsilon)h}|f_i-f_\infty|=0\Big\}$$

and the operator

$$P_I(h): A_{h,I}^{\sigma-\varepsilon}\to B_{h,I}^{\sigma-\varepsilon} \tag{3.62}$$

which is defined as the solution operator corresponding to (3.47), i.e., the operator mapping $(\tilde v_i)_{i=I}^{\infty}$ to $(\tilde u_i)_{i=I}^{\infty}$. When we put

$$(Gu)_i=h\sum_{j=I}^{i-1} a((i-j)h)\beta_{i,j}(h,-m)u_j, \tag{3.63}$$

equations (3.44) can be rewritten in the form

$$u_i=P_I(h)(Gu+\tilde v)_i,\qquad i\ge I. \tag{3.64}$$

It follows from (3.61) that $P_I(h)$ is a bounded operator. Moreover, (3.46) implies that G has small norm. Therefore, $I-P_I(h)G$ is a nonsingular operator from $A_{h,I}^{\sigma-\varepsilon}$ into $B_{h,I}^{\sigma-\varepsilon}$, and the norm of the inverse is bounded uniformly with respect to h, t_{-m} and ε. Therefore,

$$\sup_{i>0} e^{i(\sigma-\varepsilon)h}\Big|u_i-\lim_{i\to\infty} u_i\Big|+\Big|\lim_{i\to\infty} u_i\Big|\le c\|\hat v\|_{Y_h^{\sigma-\varepsilon}} \tag{3.65}$$

and, summarizing, we obtain

$$\|\hat u\|_{X_h^{\sigma-\varepsilon}}\le \text{const.}\|\hat v\|_{Y_h^{\sigma-\varepsilon}} \tag{3.66}$$

where the constant is independent of h, t_{-m} and $\varepsilon\in[0,\varepsilon_0]$, where $\varepsilon_0<\sigma$. This concludes the stability proof for $\mu=0$.

We briefly sketch the stability proof for $\mu>0$. Equation (3.19) now takes the form

$$\mu\frac{u_i-u_{i-1}}{h}=-H_i(h,-m)u_i+h\sum_{j=-m+1}^{i-1} a((i-j)h)\left(2\frac{y^3(t_i)}{y^3(t_j)}+1\right)u_j+v_i,\qquad i>-m \tag{3.67}$$

where

$$\begin{aligned} H_i(h,-m)=3\int_{-\infty}^{t_{-m}} a(t_i-s)y^2(t_i)\,ds+3h\sum_{j=-m+1}^{i-1} a((i-j)h)\frac{y^2(t_i)}{y^2(t_j)}\\ -\alpha f(t_i)y^{\alpha-1}(t_i). \end{aligned} \tag{3.68}$$

With $D(h)$ as in (3.46),

$$H_i(h,-m)=\frac{3}{D(h)}+\begin{cases} O(e^{-\sigma t_i}) & \text{as } t_i\to\infty,\\ O(h)+O(e^{\sigma t_i}) & \text{as } t_i\to-\infty. \end{cases} \tag{3.69}$$

We can therefore use a similar perturbation approach as before, i.e., for $t_i\to\infty$, (3.67) is regarded as a perturbation of the problem

$$\mu\frac{\tilde u_i-\tilde u_{i-1}}{h}=-\frac{3}{D(h)}\tilde u_i+3h\sum_{j=-m+1}^{i-1}a((i-j)h)\tilde u_j+\tilde v_i. \tag{3.70}$$

This can be rewritten as follows:

$$\tilde u_i=\gamma(h,\mu)\tilde u_{i-1}+\delta(h,\mu)h\sum_{j=-m+1}^{i-2}a((i-j)h)\tilde u_j+\tilde v_i, \tag{3.71}$$

where

$$\gamma(h,\mu)=\left(\frac{\mu}{h}+3ha(h)\right)\left(\frac{\mu}{h}+\frac{3}{D(h)}\right)^{-1},\qquad \delta(h,\mu)=3\left(\frac{\mu}{h}+\frac{3}{D(h)}\right)^{-1}.$$

We substitute

$$\tilde g^l_{i-1}=hK_l\sum_{j=-m+1}^{i-2}e^{-\lambda_l(t_{i-1}-t_j)}\tilde u_j,\qquad \tilde z_{i-1}=\sum_{l=1}^{N}e^{-\lambda_l h}\tilde g^l_{i-1}. \tag{3.72}$$

We set $\tilde w_i=(\tilde u_i,\tilde z_i,\tilde g_i^1,\cdots,\tilde g_i^N)$, $\tilde\varphi_i=(\delta(h,\mu)(\tilde v_i-v_{i-1}),0,\cdots,0)$. Then (3.70) is equivalent to the system

$$\tilde w_i=(I+F(h,\mu))\tilde w_{i-1}+\tilde\varphi_i \tag{3.73}$$

with

(3.74)

$$F(h,\mu)=\begin{pmatrix} \gamma(h,\mu)-1 & \delta(h,\mu) & 0\cdots & & 0\\ h\sum_{l=1}^{N}K_le^{-2\lambda_l h} & 0 & e^{-\lambda_1 h}(e^{-\lambda_1 h}-1) & \cdots & e^{-\lambda_N h}(e^{-\lambda_N h}-1)\\ hK_1e^{-\lambda_1 h} & 0 & e^{-\lambda_1 h}-1 & & \\ \vdots & \vdots & 0 & \ddots & 0\\ hK_Ne^{-\lambda_N h} & 0 & & & e^{-\lambda_N h}-1 \end{pmatrix}$$

As before, it can be shown that the characteristic polynomial of $(d/dh)F(0,\mu)$ has the same roots as (2.3), except for the fact that 0 is a double rather than a simple root. Moreover, 0 is an exact eigenvalue of $F(h,\mu)$.

A proof analogous to the one for $\mu=0$ shows the stability for μ fixed and sufficiently small h. For the limit $\mu\to0$, a different argument is needed. When we substitute in (3.73)

$$\tilde u_i+\frac{\delta(h,\mu)}{\gamma(h,\mu)-1}\tilde z_i=\tilde p_i,\qquad \tilde z_i=\tilde q_i, \tag{3.75}$$

we obtain a system of difference equations of the form (3.73) with $F(h,\mu)$ substituted by a matrix of the following form

$$\tilde{F}(h,\mu)=\begin{pmatrix} \gamma(h,\mu)-1 & O(h) \\ O(h) & h\dfrac{d}{dh}A(0,0)+O(h^2) \end{pmatrix}. \tag{3.76}$$

Moreover, an estimate of the form $-1\le\gamma(h,\mu)-1\le-\omega\min(C_1,C_2h/\mu)$ holds where $\omega>0$. Thus, for small μ, $|\gamma-1|\gg h$. It is easy to conclude from this that there is a coordinate transformation close to the identity which transforms $\tilde{F}$ to the form

$$\tilde{\tilde{F}}=\begin{pmatrix} \gamma(h,\mu)-1+O(h) & 0 \\ 0 & h\dfrac{d}{dh}A(0,0)+o(h) \end{pmatrix}. \tag{3.77}$$

Stability follows from the above estimate for $\gamma-1$ and an analysis of the eigenvalues of $(d/dh)A(0,0)$ given by (3.39). For $t_i\to-\infty$ a similar argument holds, but $D(h)$ is to be replaced by a different constant $\tilde{D}(h,t_{-m})$. From these considerations we see that

$$\left\|L_{h,m}^{-1}\right\|_{Y_h^{\sigma-\varepsilon}\to X_h^{\sigma-\varepsilon}}\le\text{const.} \tag{3.78}$$

with a constant independent of h, $-m,\mu$ and $\varepsilon\in[0,\varepsilon_0]$, $\varepsilon_0<\sigma$.

It is practically important to assure stability not just for h sufficiently small, but also for arbitrary h. Recall that linearizing $e^{-\lambda_l h}-1$ with respect to h is only justified if $h\ll 1/\lambda_l$. The matrix $F(h,\mu)$ for arbitrary h has the same form as for h small, if the following substitutions are made: $\lambda_i\to(e^{-\lambda_i h}-1)/h$, $K_i\to K_i e^{-2\lambda_i h}$, $\mu\to\mu+3h/D(h)$. If f has compact support, this is sufficient to ensure stability. If the support of f is not compact, stability for arbitrary h can be assured if the following modification is made: In (3.3) the integral $\int_{-\infty}^{t_{-m}}a(t_i-s)\,ds$ is replaced by $h\sum_{j=-\infty}^{-m}a(t_i-t_j)$. With this modification the term $O(h)$ in (3.69) vanishes, and thus the matrix for the linearized problem is asymptotically equal to $F(h,\mu)$ both for $t_i\to\infty$ and $t_i\to-\infty$. In order to apply Keller's [3] nonlinear stability concept, it is further necessary to show that the Fréchet derivatives $D_{\hat{z}}F_{h,m}$ are uniformly Lipschitz continuous in a sphere

$$S_K=\left\{\hat{z}\in X_h^{\sigma-\varepsilon}\,\middle|\,\left\|\hat{z}-(y(t_i)-1)_{i=-\infty}^{\infty}\right\|<K\right\}.$$

This follows from a fairly trivial calculation, which we do not present here.

Using the fact that the global error $(y(t_i)-y_i)_{i=-\infty}^{\infty}$ is estimated by a constant times the bound for the local error (3.12), we obtain the following theorem:

THEOREM 3.1. *The discretization scheme* (3.3), (3.4) *has a unique solution for all* $f\in Y^{\sigma,1}$ *which either have small norm or vanish identically for* $t\ge t_0$ *for some finite* t_0, *where* σ *is as in Theorem* 2.3. *This solution* $\hat{y}=(y_i)_{i=-\infty}^{\infty}$ *can be calculated by the Newton procedure which is second order convergent from a sphere of starting values which does not shrink to* $\varnothing$ *as* $h\to0$, $t_{-m}\to-\infty$, $i\to0$, *and the convergence estimate*

$$\left\|(y_i-y(t_i))_{i=-\infty}^{\infty}\right\|_{X_h^{\sigma-\varepsilon}}\le\text{const.}\left(h+o\left(e^{-\varepsilon(t_{-m})}\right)\right) \tag{3.79}$$

holds for h sufficiently small and $|t_{-m}|$ *sufficiently large. The constant is independent of* h, t_{-m}, $\mu\in[0,\infty]$, $\varepsilon\in[0,\varepsilon]$, $\varepsilon_0<\sigma$.

This implies that the Newton procedure for the solution of (3.6) can be safely applied, that the $(y_i)_{i=-\infty}^{\infty}$ do not exhibit boundary-layer-like behavior, and that

$$|y_i-y(t_i)|\le\text{const.}\,e^{(\sigma-\varepsilon)t_i}\left(h+o\left(e^{-\varepsilon|t_{-m}|}\right)\right),\qquad t_i\le0, \tag{3.80}$$

$$\left|\lim_{i\to\infty}y_i-y(\infty)\right|\le\text{const.}\left(h+o\left(e^{-\varepsilon|t_{-m}|}\right)\right), \tag{3.81}$$

$$(3.82)\quad \left|\left(y_i-\lim_{i\to\infty} y_i\right)-\left(y(t_i)-y(\infty)\right)\right|\leq \text{const.}\, e^{-(\sigma-\varepsilon)t_i}\left(h+o(e^{-\varepsilon|t_{-m}|})\right),\qquad t_i\geq 0,$$

and the order of convergence is independent of $\mu\in[0,\infty]$.

Obviously, if f is only supported on $[T_1,T_2]$, then the term $O(e^{-\varepsilon|t_{-m}|})$ disappears from the error estimate if $t_{-m}<T_1$.

The discretization we used was derived from the integral equation. In §2 we transformed to a system of ordinary differential equations. In fact, up to terms of order $O(h)$, our discretization method corresponds to a discretization scheme for the ODE system (2.1). Namely, if we put

$$g_{l,i}=h\sum_{j=-m+1}^{i} K_l e^{-\lambda_l(t_i-t_j)}\frac{1}{y_j^2},\qquad g_{l,-m}=0,$$

$$h_{l,i}=h\sum_{j=-m+1}^{i} K_l e^{-\lambda_l(t_i-t_j)}y_j,\qquad h_{l,-m}=0,$$

our discretized equation reads as follows:

$$(3.83)\quad \text{(a)}\qquad -\mu\left(\frac{y_i-y_{i-1}}{h}\right)=\sum_{l=1}^{N} e^{-\lambda_l h}(g_{l,i-1}y_i^3-h_{l,i-1})-f_i y_i^{\alpha}+\int_{-\infty}^{t_{-m}} a(t_i-s)\,ds\,(y_i^3-1),$$

$$\text{(b)}\qquad \frac{g_{l,i}-g_{l,i-1}}{h}=\frac{K_l}{y_i^2}+\frac{(e^{-\lambda_l h}-1)}{h}g_{l,i-1},$$

$$\text{(c)}\qquad \frac{h_{l,i}-h_{l,i-1}}{h}=K_l y_i+\frac{(e^{-\lambda_l h}-1)}{h}h_{l,i-1}.$$

By calculating $g_{l,i}$, $h_{l,i}$ for $l=1(1)N$ from (3.83)(b),(c) and by inserting these quantities into (3.83)(a), we obtain an equation (in each time step) of the form (3.6). Theorem 3.1 now implies that the root of this equation can be safely obtained by the Newton procedure which is second order accurate from a sphere of starting values whose radius is independent of h, t_{-m}, $\mu\in[0,\infty)$ and $i>-m$.

This provides us with a very efficient method of solving the approximating problems, and Theorem 3.1 makes sure that the qualitative properties of the solution of (1.1) carry over to the approximate solutions.

The exponential decay of the solution encourages one to attempt using variable mesh sizes of the form

$$(3.84)\qquad \hat{h}_i=\hat{h}e^{-\sigma|t_i|}.$$

It can be expected that convergence of the order of one in $\hat{h}$ (i.e., the estimate would be $O(\hat{h})+o(e^{-\sigma|t_{-m}|})$) would follow in l^∞, but the exponential decay property of the approximate solution would be lost. In the case of boundary value problems for ordinary differential equations on infinite intervals, this has been shown in [7].

A further problem that should be mentioned is, which higher order discretization schemes could be employed. It is fairly clear from our analysis that polynomial collocation methods using Radau points (see [12]) would recover the exponential decay property and the uniform convergence as $\mu\to 0$.

4. Numerical results. For the computations we used the kernel $a(u)=\sum_{i=1}^{8} K_i e^{-\lambda_i u}$ with the constants K_i and λ_i shown in Table 1.

TABLE 1

i	$\lambda_i(\text{sec}^{-1})$	$K_i(\text{Nm}^{-2}\text{sec}^{-1})$
1	10^{-3}	1×10^{-3}
2	10^{-2}	1.8×10^{0}
3	10^{-1}	1.89×10^{2}
4	1	9.8×10^{3}
5	10	2.67×10^{5}
6	10^{2}	5.86×10^{6}
7	10^{3}	9.48×10^{7}
8	10^{4}	1.29×10^{9}

These numbers were obtained by Laun [4] from an experimental fit for a polyethylene melt at 150°C, which he calls "Melt 1".

The parameter μ is physically identified as three times the Newtonian contribution to the viscosity. Experimental values are not available, and theoretically μ is either a solvent viscosity (for polymer solutions) or it results from fractions of low molecular weight (for melts). The value of μ has to be compared to the viscosity resulting from the memory, which, for constant shear rate, is given by $\sum_{i=1}^{8} K_i\lambda_i^{-2}\approx 50\,000\ \text{Nm}^{-2}\text{sec}$. One would expect μ to influence the solution significantly only if it exceeds this value. This is verified by our computations. In the plots (Figs. 2–15), the scale for y is on the left, the scale for f is on the right. y is measured in multiples of the length (for the filament, $\alpha=2$), or, respectively, the thickness (for the sheet, $\alpha=\frac{1}{2}$) at $t=-\infty$. f denotes the force acting on the ends of the filament or the edges of the sheet divided by the cross-sectional area in the undeformed state at $t=-\infty$; f is expressed in N/m^2. The time is measured in seconds. f is always plotted by dashed lines, y by full lines.

All plots except Figs. 8, 9 were made for $\alpha=2$, the case of the filament. In Figs. 2–12 (except 7), the force f is of the form

$$f(t)=\begin{cases}0, & |t|\geq a_2,\\ f_{\max}\exp\left\{a_0^2-\dfrac{a_1^2}{a_2^2-t^2}\right\}, & |t|<a_2,\end{cases}$$

with $a_0^2-a_1^2/a_2^2=0$. Such an f is in $C^\infty(\mathbb{R},\mathbb{R})$ and has the compact support $[-a_2,a_2]$.

The parameter μ is zero in Figs. 2–10. In Figs. 2–6 we have chosen various values of $f_{\max}$, a_0 and a_1, as can be seen from the diagrams. The calculations were done for larger time intervals than the plots, thus yielding approximations for $y(\infty)$. For Figs. 2–6, the approximate values of $y(\infty)$ are as follows [2]:

Fig.	2	3	4	5	6
$y(\infty)$	1.07	1.15	1.11	1.26	1.13

[2] In Fig. 6 supp$f$ is different from that in the previous ones.

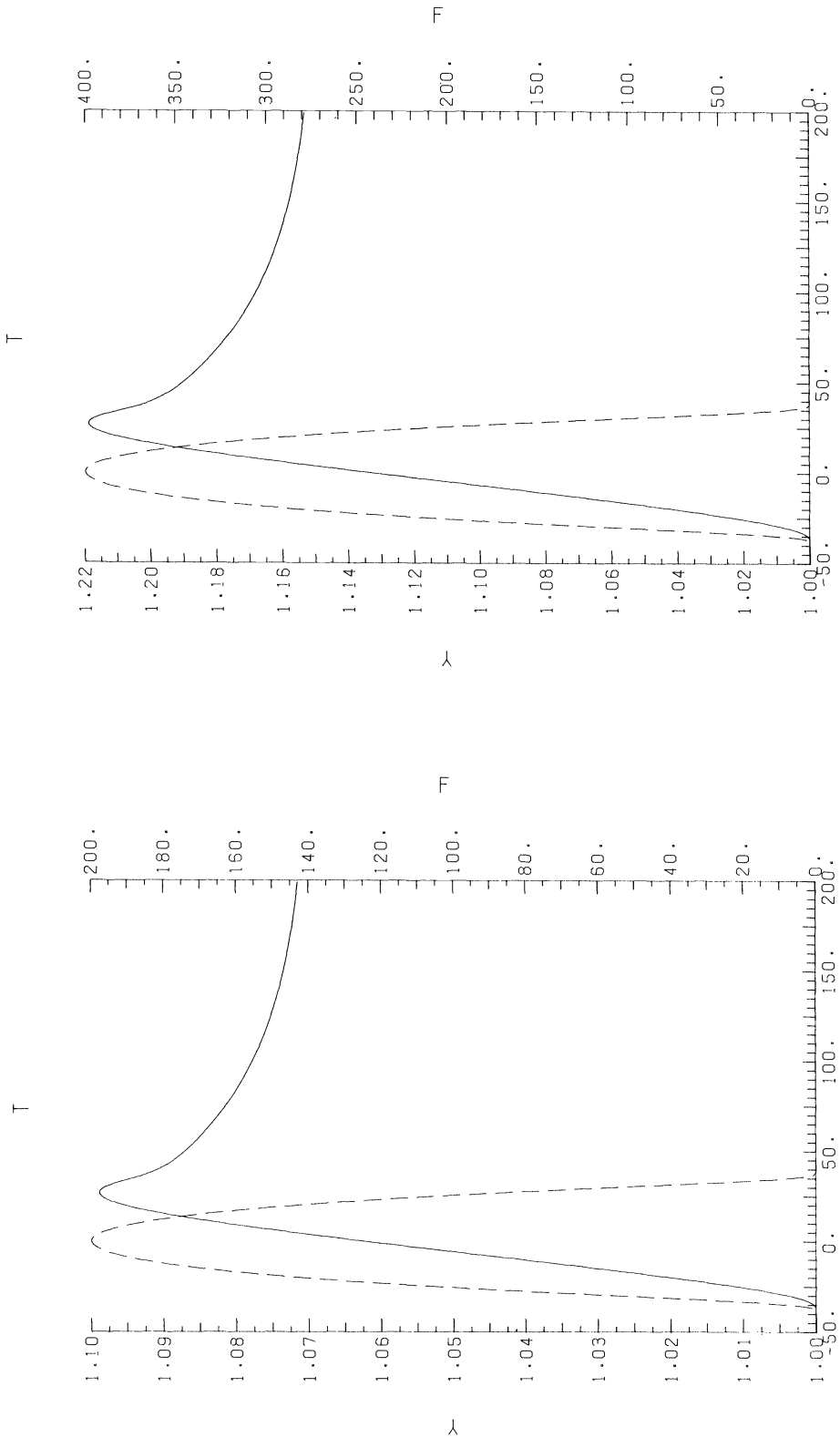

T
F
400.
350.
300.
250.
200.
150.
100.
50.
0.
-50.
1.22
1.20
1.18
1.16
1.14
1.12
1.10
1.08
1.06
1.04
1.02
1.00
0.
50.
100.
150.
200.
T
F
T
F
200.
180.
160.
140.
120.
100.
80.
60.
40.
20.
0.
-50.
1.10
1.09
1.08
1.07
1.06
1.05
1.04
1.03
1.02
1.01
1.00
0.
50.
100.
150.
200.
T
F

FIG. 3

FIG. 2

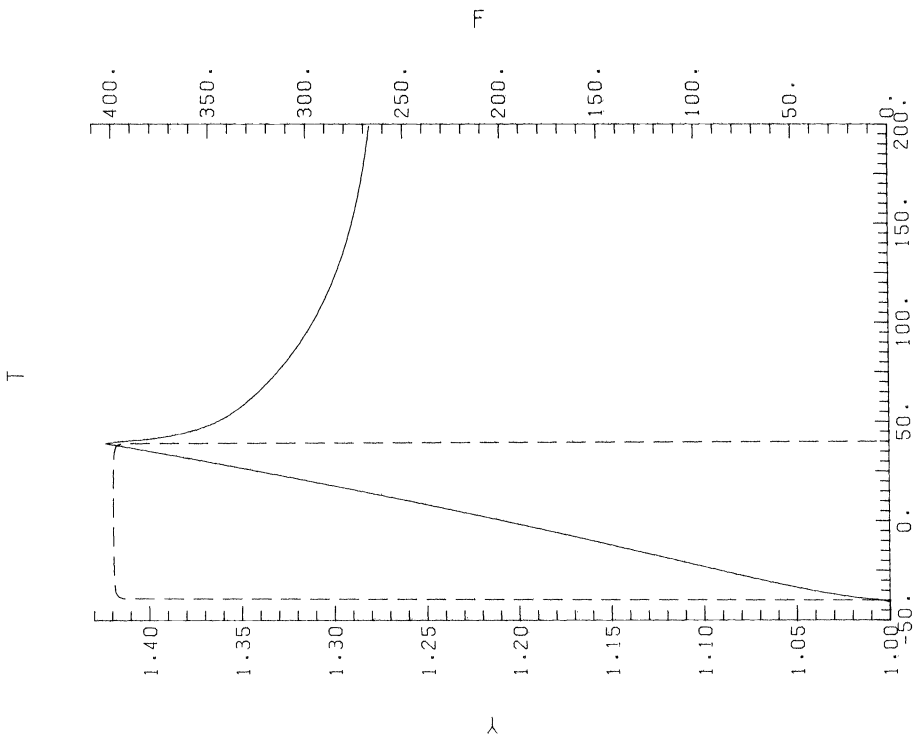

T
F
400.
350.
300.
250.
200.
150.
100.
50.
0.
200.
150.
100.
50.
0.
-50.
T
1.40
1.35
1.30
1.25
1.20
1.15
1.10
1.05
1.00
Y

FIG. 5

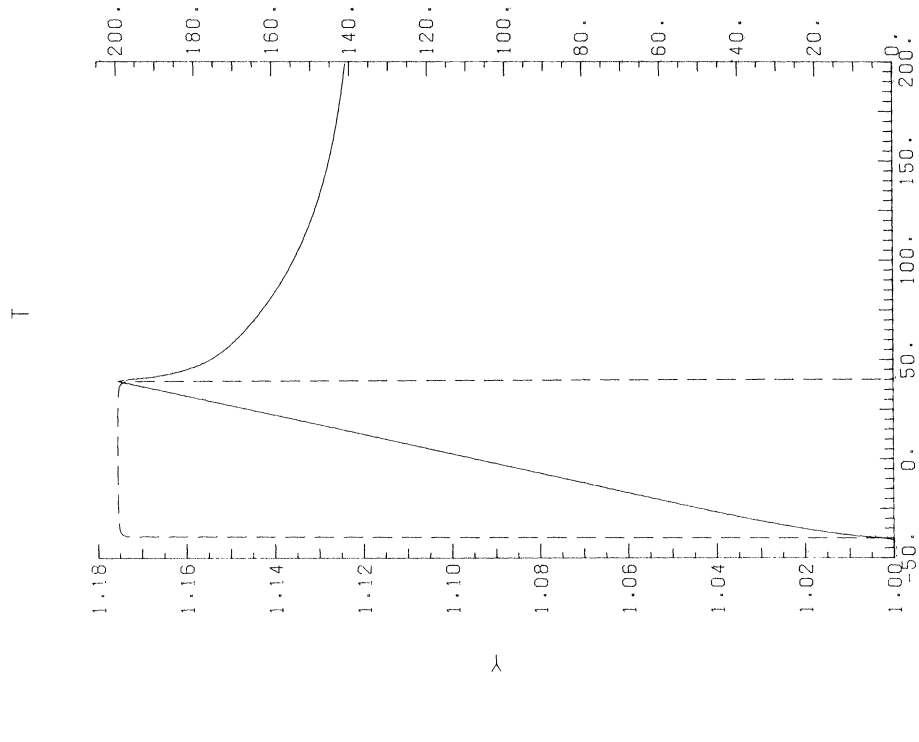

T
F
200.
180.
160.
140.
120.
100.
80.
60.
40.
20.
0.
200.
150.
100.
50.
0.
-50.
T
1.18
1.16
1.14
1.12
1.10
1.08
1.06
1.04
1.02
1.00
Y

FIG. 4

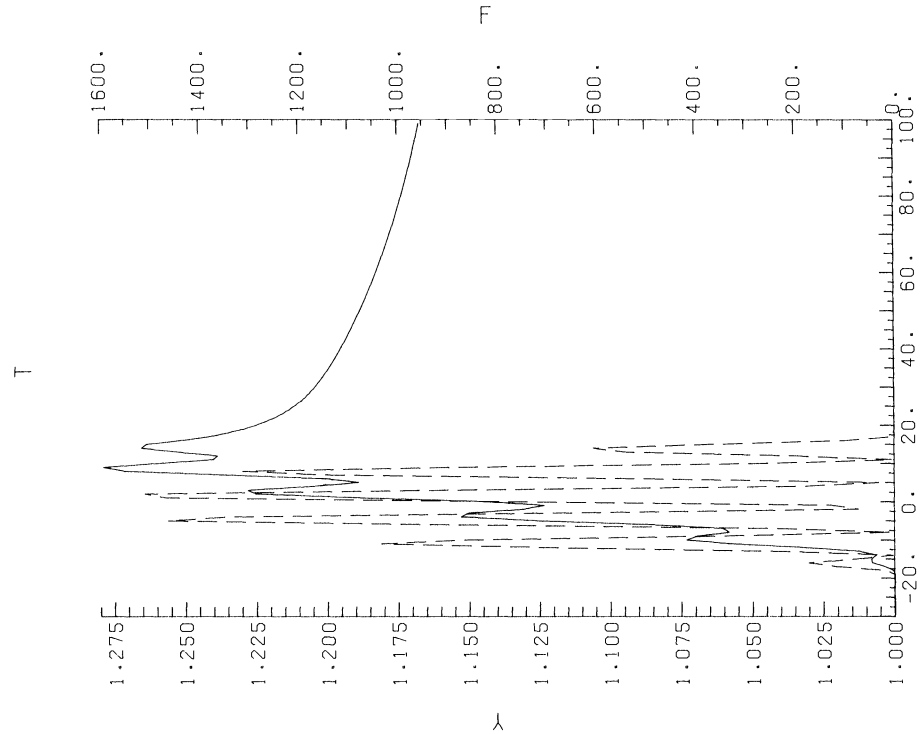

T
F
1600.
1400.
1200.
1000.
800.
600.
400.
200.
0.
-20. 0. 20. 40. 60. 80. 100.
1.275
1.250
1.225
1.200
1.175
1.150
1.125
1.100
1.075
1.050
1.025
1.000
Y
T

FIG. 7

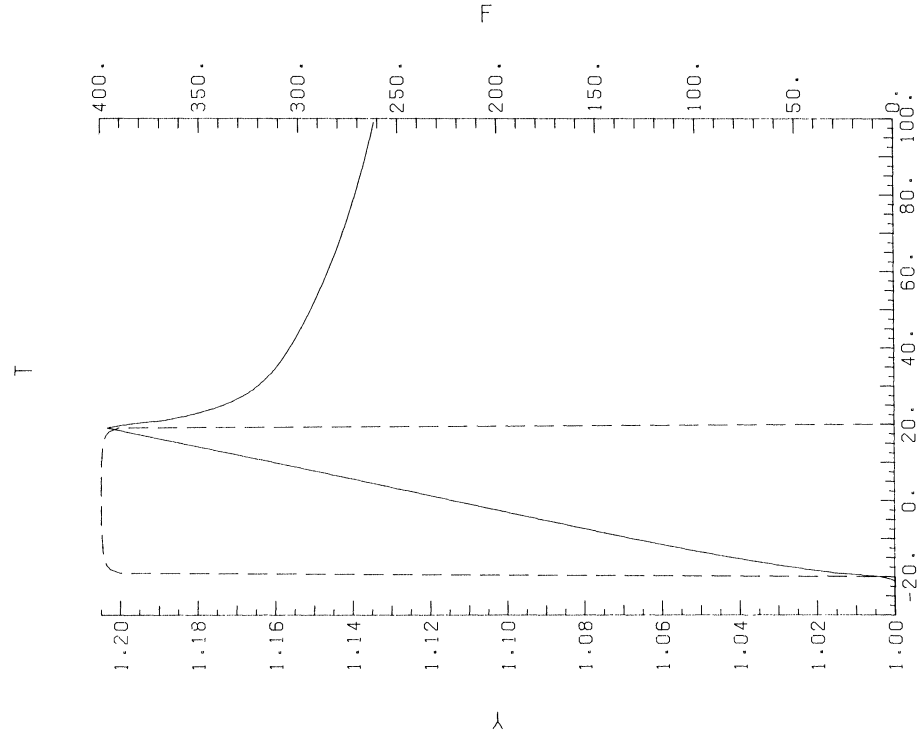

T
F
400.
350.
300.
250.
200.
150.
100.
50.
0.
-20. 0. 20. 40. 60. 80. 100.
1.20
1.18
1.16
1.14
1.12
1.10
1.08
1.06
1.04
1.02
1.00
Y
T

FIG. 6

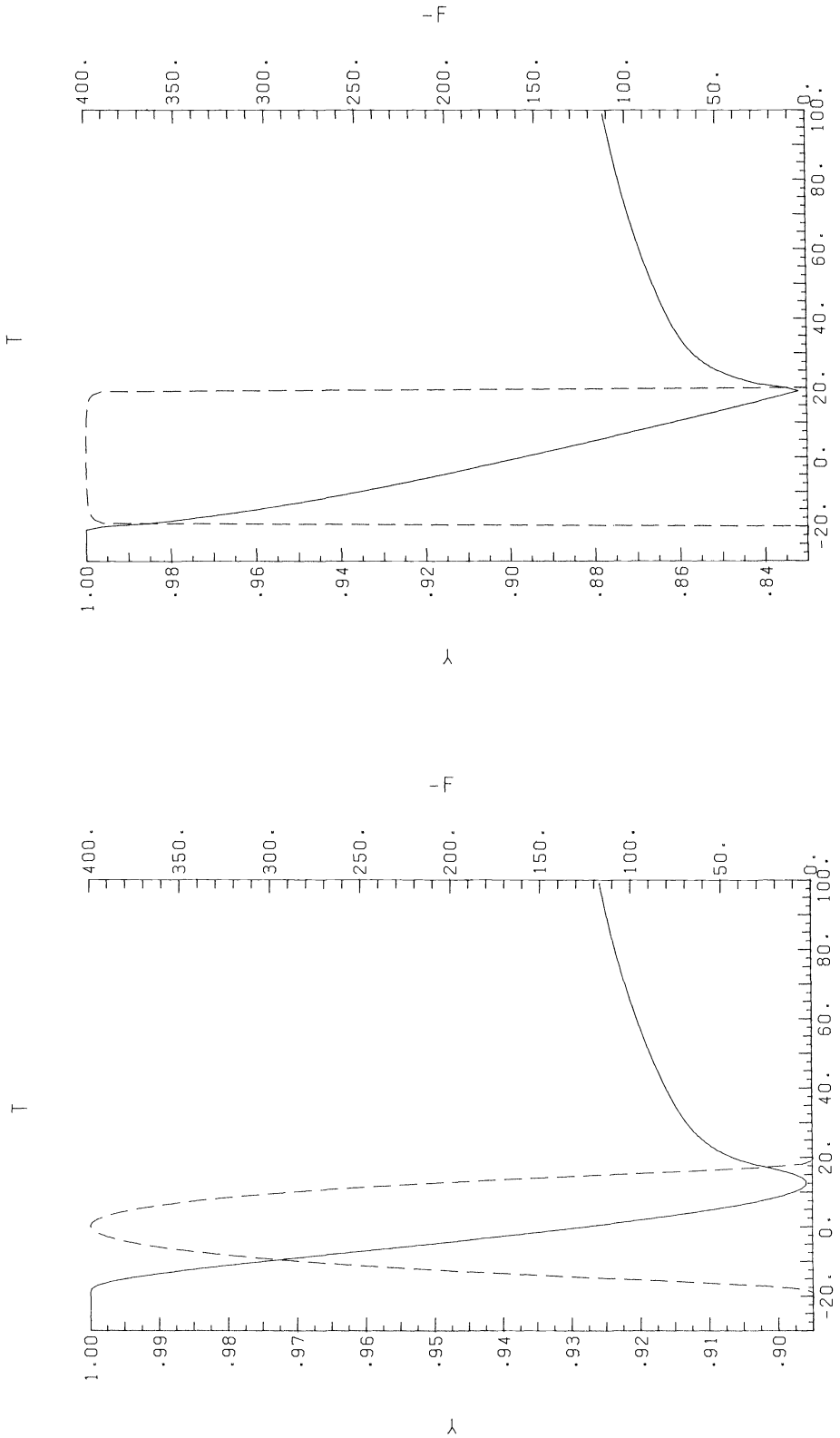

-F
400.
350.
300.
250.
200.
150.
100.
50.
0.
100.
80.
60.
40.
20.
0.
-20.
T
1.00
.98
.96
.94
.92
.90
.88
.86
.84
Y
T
T

FIG. 9

-F
400.
350.
300.
250.
200.
150.
100.
50.
0.
100.
80.
60.
40.
20.
0.
-20.
T
1.00
.99
.98
.97
.96
.95
.94
.93
.92
.91
.90
Y
T
T

FIG. 8

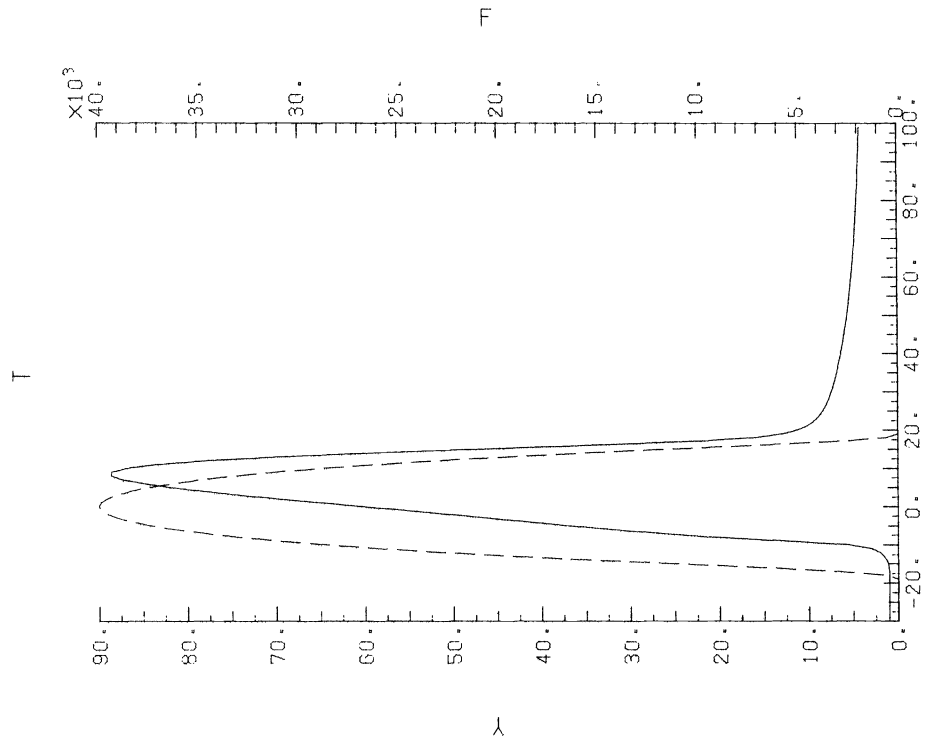

T
F
×10^3
40.
35.
30.
25.
20.
15.
10.
5.
0.
-20.
0.
20.
40.
60.
80.
100.
T
90.
80.
70.
60.
50.
40.
30.
20.
10.
0.
λ

FIG. 11

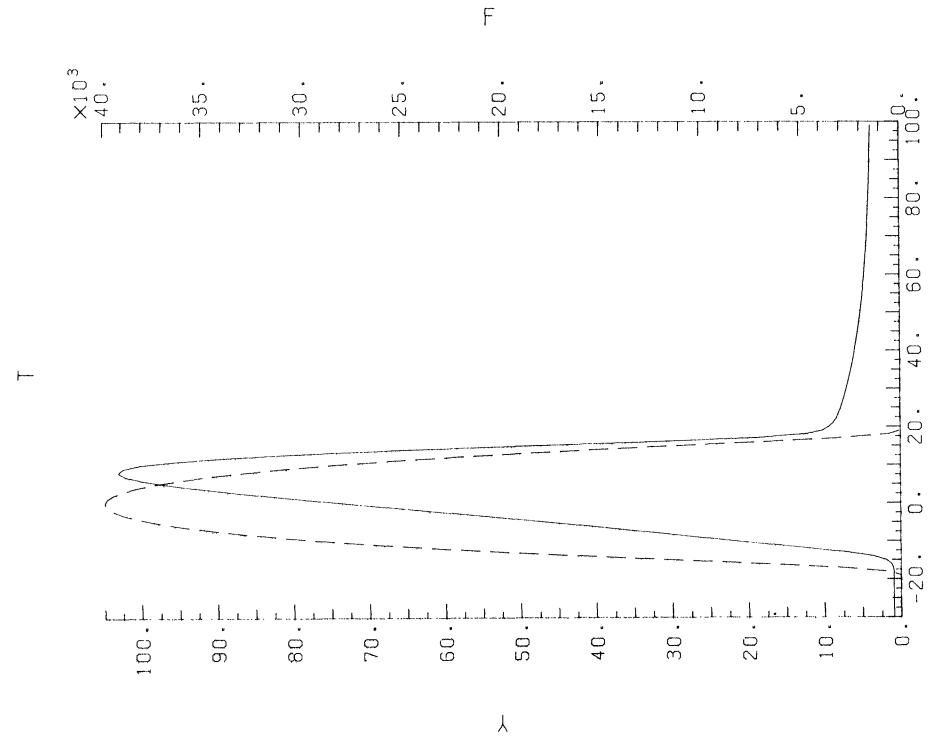

T
F
×10^3
40.
35.
30.
25.
20.
15.
10.
5.
0.
-20.
0.
20.
40.
60.
80.
100.
T
100.
90.
80.
70.
60.
50.
40.
30.
20.
10.
0.
λ

FIG. 10

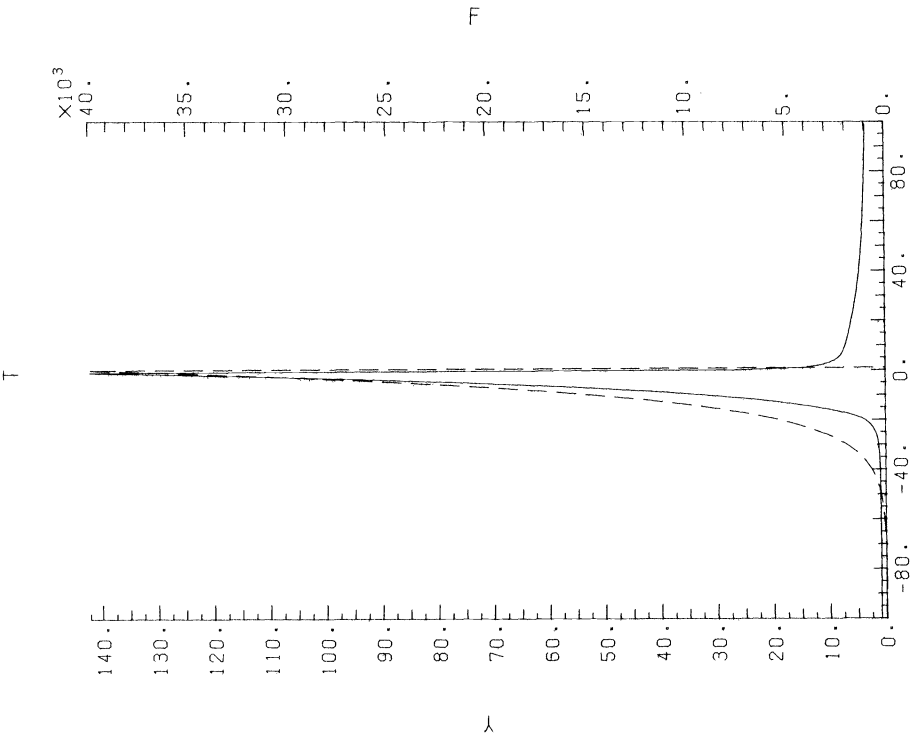


FIG. 13. $\mu=0$.

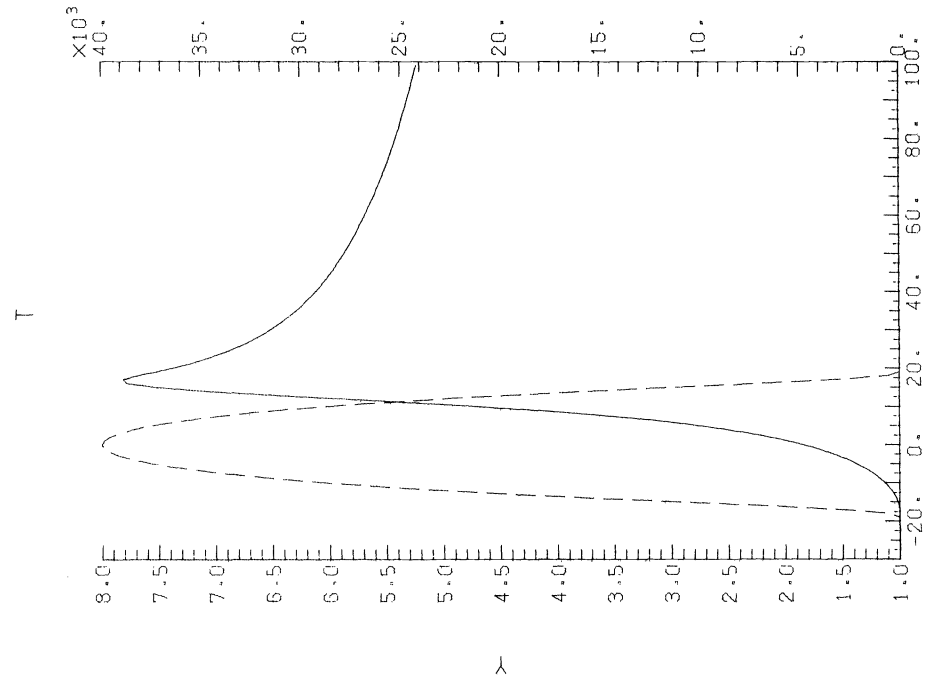


FIG. 12

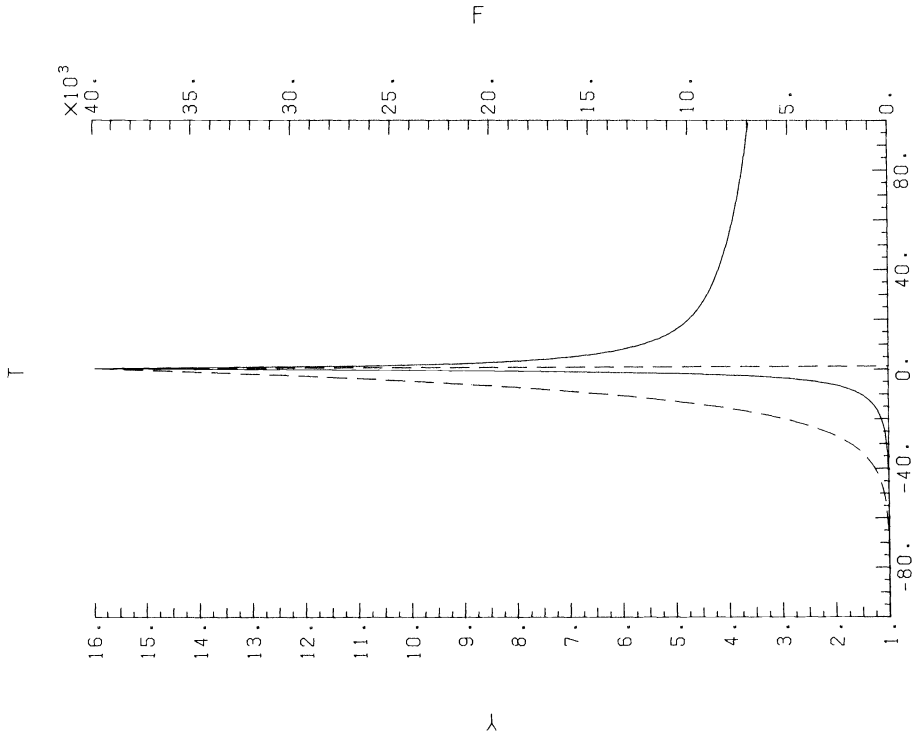


FIG. 15. $\mu = 35 \times 10^4$.

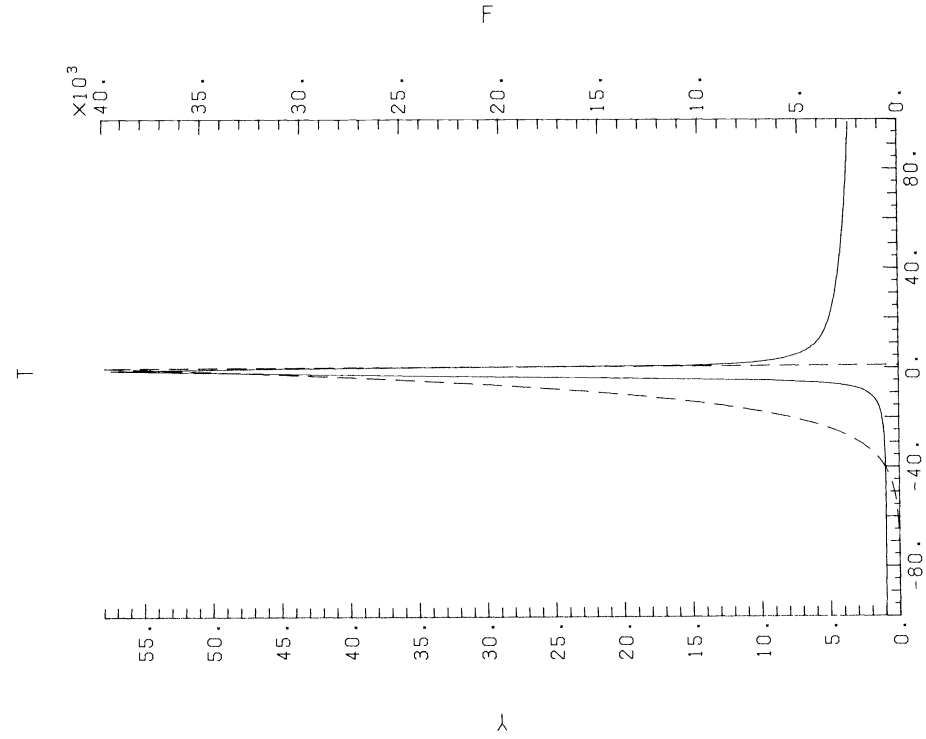


FIG. 14. $\mu = 20 \times 10^4$.

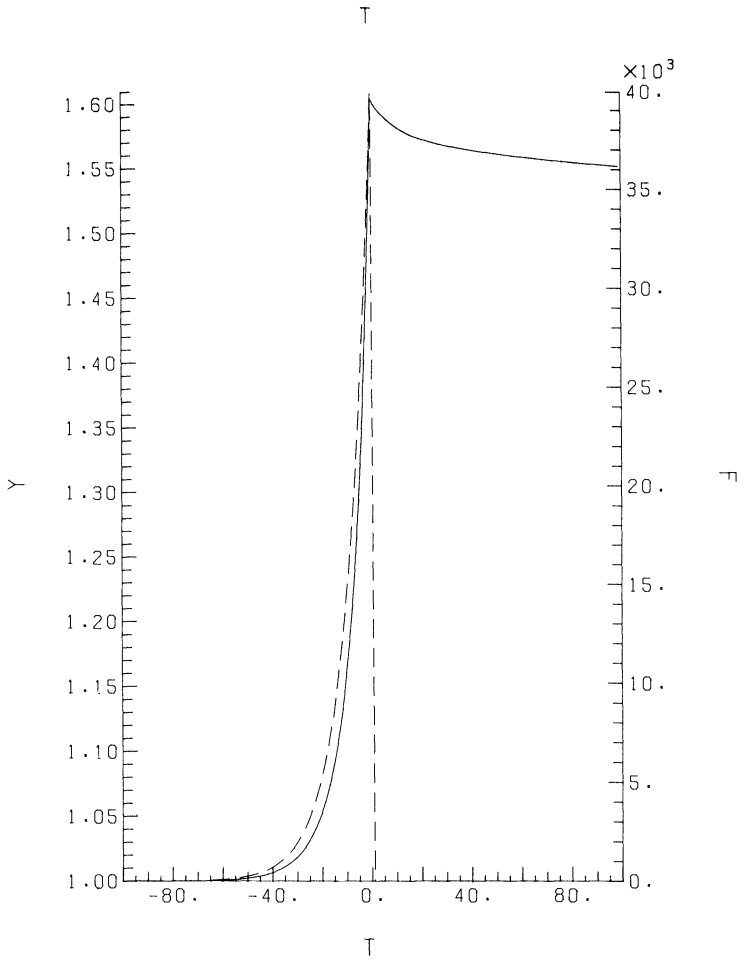


FIG. 16. $\mu = 10^6$.

These numbers indicate that, roughly speaking, the value of $y(\infty) - 1$ is proportional to $\int_{-\infty}^{\infty} f(t)\,dt$. This would in fact be exact for the linearized equation.

In Fig. 7 an oscillating force was chosen. It is observed that the solution y "follows" the oscillations with a certain delay.

Figures 8 and 9 illustrate the case of the sheet ($\alpha = \frac{1}{2}$). Here $-f$ is plotted rather than f. The results are qualitatively similar to those in Figs. 2–6, but now we have $y < 1$ instead of $y > 1$.

In Figs. 10–16, we have again $\alpha = 2$. In Figs. 10–12, we have chosen the same $f(f_{\max} = 40\,000,\ a_0 = 1,\ a_1 = a_2 = 20)$ and computed solutions for different values of μ:

μ	$y_{\max}$	$y(\infty)$
0	103	3.3
10^5	89	3.4
10^6	7.7	4.5

For $\mu \leq 10\,000$, no significant change was observed. For larger μ, the effect on the maximal elongation seems to be more pronounced than the effect on the final length. Recalling the fact that $\mu = 10\,000$ would correspond to a viscosity 3×10^6 as large as that of water, it seems conceivable that for fluids like "Melt 1" μ can be neglected.

The numbers for $y(\infty)$ are interesting in comparison with the results of Lodge, McLeod and Nohel [6]. They showed that $y(\infty)$ increases with μ, if the history of y for $t<0$ is kept fixed. Our numbers show the same tendency; eventually, however, $y(\infty)$ has to decrease, since for $\mu=\infty$ we have $y=$const. and thus $y(\infty)=1$. We see from this that, for fixed f, $y(\infty)$ is not a monotone function of μ.

For Figs. 13–16, a discontinuous force given by

$$f(t)=\begin{cases}0, & t\geq 0,\\ 40\,000\exp\left(\dfrac{t}{10}\right), & t<0,\end{cases}$$

was used. Since in this case the filament recovers freely for $t>0$, we are studying the same situation as Lodge, McLeod and Nohel [6], but we prescribe the force rather than the history of y for $t<0$. By considering the intervals $t<0$ and $t>0$ separately, we can easily modify the existence and convergence theory of the previous sections for the present case. However, the solution does not depend continuously on μ in the L^∞-norm as $\mu\to 0$. This is because for $\mu=0$ the solution is discontinuous at $t=0$. Table 2 illustrates the dependence of the maximal elongation on μ.

TABLE 2

μ	$y_{\max}$
0	140
2×10^5	58
3.5×10^5	16
10^6	1.6

Acknowledgment. The authors would like to thank Professors John Nohel and Arthur Lodge for helpful discussions.

REFERENCES

[1] H. BRUNNER AND J. D. LAMBERT, *Stability of numerical methods for Volterra-integrodifferential equations*, Computing, 12 (1974), pp. 75–84.

[2] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin-Heidelberg-New York, 1966.

[3] H. B. KELLER, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29 (1975), pp. 464–474.

[4] H. M. LAUN, *Description of the non-linear shear behaviour of a low density polyethylene melt by means of an experimentally determined strain dependent memory function*, Rheol. Acta, 17 (1978), pp. 1–15.

[5] A. S. LODGE, *Body Tensor Fields in Continuum Mechanics*, Academic Press, New York-San Francisco-London, 1974.

[6] 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 Sect. A, 80 (1978), pp. 99–137.

[7] P. A. MARKOWICH AND C. RINGHOFER, *Boundary value problems on long intervals*, MRC TSR 2205, Mathematics Research Center, Univ. of Wisconsin, Madison, 1981; (submitted to Math. Comp.)

[8] O. NEVANLINNA, *Numerical solution of a singularly perturbed nonlinear Volterra equation*, MRC TSR 1881, Univ. of Wisconsin, Madison, 1978.

[9] M. RENARDY, *A quasilinear parabolic equation describing the elongation of thin filaments of polymeric liquids*, MRC TSR 2183, Univ. of Wisconsin, Madison, 1981; this Journal, 13 (1982), pp. 226–238.

[10] _____, *Bifurcation from rotating waves*, Arch. Rational Mech. Anal., to appear.

[11] _____, *Bifurcation of singular and transient solutions: Spatially nonperiodic patterns for chemical reaction models in infinitely extended domains*, in Recent Contributions to Nonlinear Partial Differential Equations, H. Berestycki and H. Brezis, eds., Pitman, London-San Francisco-Melbourne, 1981.

[12] R. WEISS, *The application of implicit Runge–Kutta and collocation methods to boundary value problems*, Math. Comp., 28 (1974), pp. 449–464.

[13] L. CESARI, *Asymptotic Behaviour and Stability Properties of Ordinary Differential Equations*, Springer-Verlag, Berlin, 1971.

[14] W. V. CHANG, R. BLOCH AND N. W. TSCHOEGL, *On the theory of viscoelastic behaviour of soft polymers in moderately large deformations*, Rheol. Acta, 15 (1976), pp. 367–378.

[15] M. W. JOHNSON AND D. SEGALMAN, *A model for viscoelastic fluid behaviour which allows non-affine deformation*, J. Non-Newt. Fluid Mech., 2 (1977), pp. 255–270.

[16] R. W. OGDEN, *Large deformations isotropic elasticity—On the correlation of theory and experiment in incompressible rubberlike solids*, Proc. Roy. Soc. London, Ser. A, 326 (1972), pp. 565–584.

[17] C. J. S. PETRIE, *On stretching Maxwell models*, J. Non-Newt. Fluid Mech., 2 (1977), pp. 221–253.

[18] _____, *Elongational Flows*, Pitman, London, 1979.

[19] B. R. SETH, in Second Order Effects in Elasticity, Plasticity and Fluid Mechanics, M. Reiner and D. Abir, eds., Pergamon, New York, 1964.

[20] N. PHAN THIEN AND R. I. TANNER, *A new constitutive equation derived from network theory*, J. Non-Newt. Fluid Mech., 2 (1977), pp. 353–365.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0007 $01.25/0

LINEAR FUNCTIONAL DIFFERENTIAL EQUATIONS AS SEMIGROUPS ON PRODUCT SPACES*

JOHN A. BURNS[†‡], TERRY L. HERDMAN[†§] AND HARLAN W. STECH[†¶]

Abstract. In this paper we consider the well-posedness of linear functional differential equations on product spaces. Let L and D be linear $\mathbb{R}^n$-valued functions with domains $\mathcal{D}(L)$ and $\mathcal{D}(D)$ subspaces of the Lebesgue measurable $\mathbb{R}^n$-valued functions on $[-r,0]$ and such that $W^{1,p}([-r,0];\mathbb{R}^n)\subseteq\mathcal{D}(L)\cap\mathcal{D}(D)$. Under weak conditions on D and L we establish the equivalence between generalized solutions to the functional differential equation

$$\frac{d}{dt}Dx_t=Lx_t+f(t)$$

and mild solutions to the Cauchy problem in $\mathbb{R}^n\times L_p([-r,0];\mathbb{R}^n)$

$$\dot{z}(t)=\mathcal{A}z(t)+(f(t),0),$$

where $\mathcal{A}$ is the operator defined on

$$\mathcal{D}(\mathcal{A})=\{(\eta,\varphi)\in\mathbb{R}^n\times L_p([-r,0];\mathbb{R}^n)/\varphi\in W^{1,p}([-r,0];\mathbb{R}^n), D\varphi=\eta\},$$

by

$$\mathcal{A}(\eta,\varphi)=(L\varphi,\dot{\varphi}).$$

The results are applicable to neutral functional differential equations and certain singular integral equations.

1. Introduction. During the past few years it has been recognized that product spaces provide appropriate state spaces for the investigation of certain problems involving control systems governed by retarded functional differential equations (RFDEs). These spaces have been used by a number of authors (see [2], [26] for a survey of the literature), and are especially well suited for the investigation of approximation techniques for identification and optimal control of RFDE systems (see [2], [3], [11], [13], [21], [22], [24]). One reason that the product spaces $\mathbb{R}^n\times L_p$ are particularly useful state spaces for RFDEs is that it can be shown that the original RFDE system can be *equivalently* formulated as a linear (ordinary differential) control system $\dot{z}=\mathcal{A}z+\beta u$ in $\mathbb{R}^n\times L_p$ (see [2], [3] and [21]). One is then able to make use of various approximation results for well-posed Cauchy problems (i.e., Trotter–Kato theorems [27], [29], finite-difference methods [17], [28], etc.) to develop computational algorithms and establish convergence of numerical schemes for RFDE systems. Moreover, the product space structure has been used to study questions of stability, controllability and observability for retarded systems [12], and certain differential-boundary operators [7].

In the present paper we consider the well-posedness of a large class of linear functional differential equations, including neutral functional differential equations (NFDEs) and certain singular integral equations. In particular, we show that these equations may be formulated as equivalent linear dynamical systems on product spaces.

---

* Received by the editors December 15, 1980, and in revised form November 12, 1981.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡ The work of this author was supported in part by the Army Research Office under contract DAAG-29-80-C-0126, the Air Force Flight Dynamics Laboratory under grant AFOSR-80-0068 and the National Science Foundation under grant ECS-8109245.

§ The work of this author was supported in part by the Air Force Flight Dynamics Laboratory under grant AFOSR-80-0068.

¶ The work of this author was supported in part by the National Science Foundation under grant MCS 81-02420 and the Army Research Office under contract DAAG-29-80-C-0126.

Throughout the paper, the positive integer n, the real number $p \in [1, +\infty)$ and the delay $r \in (0, +\infty)$ are assumed fixed. We shall use $L_p(a,b) = L_p([a,b]; \mathbb{R}^n)$ to denote the customary Lebesgue spaces of $\mathbb{R}^n$-valued "functions" on $[a,b]$ whose components are integrable when raised to the pth power. The usual Banach space $C([a,b]; \mathbb{R}^n)$ of continuous $\mathbb{R}^n$-valued functions on $[a,b]$ will be denoted by $C(a,b)$ and similarly the Sobolev spaces (see [1]) $W^{1,p}([a,b]; \mathbb{R}^n)$ will be denoted by $W^{1,p}(a,b)$. Whenever $[a,b] = [-r,0]$ we shall simply write L_p, C and $W^{1,p}$ for $L_p(-r,0)$, $C(-r,0)$ and $W^{1,p}(-r,0)$, respectively. We shall use $\|\cdot\|_X$ to denote the norm on the normed linear space X. However, we will use the same symbol $\|\cdot\|$ to denote any one of several norms when it is clear from the context which norm is intended. The space of bounded linear operators from X to the normed linear space Y will be represented by $\mathcal{B}(X, Y)$. For a linear operator $\mathcal{A}$ we use the standard notation: $\mathcal{D}(\mathcal{A})$, $\mathcal{R}(\mathcal{A})$, $\mathcal{N}(\mathcal{A})$ for the domain, range and null space of $\mathcal{A}$, respectively, and $\rho(\mathcal{A})$ will denote the resolvent set of $\mathcal{A}$. If $x: [-r, a) \to \mathbb{R}^n$ for some $0 < a \leq +\infty$, then we define $x_t: [-r, 0] \to \mathbb{R}^n$ for $0 \leq t < a$ by $x_t(s) = x(t+s)$.

While we shall not attempt to give a detailed discussion of the literature concerning the use of product spaces as state spaces for hereditary systems (see [2], [6], [8], [10]), a few comments are needed to put our presentation in this paper in prospective for the reader. In 1969, Borisovič and Turbabin [4] considered the RFDE

$$\dot{x}(t) = Lx_t + f(t) \tag{1.1}$$

with initial data

$$x(0) = \eta, \qquad x_0 = \varphi, \tag{1.2}$$

where L is a linear $\mathbb{R}^n$-valued operator, $(\eta, \varphi) \in \mathbb{R}^n \times L_2$ and $f(\cdot)$ was locally integrable on $[0, +\infty)$. Under strong assumptions on L, they showed that $\mathcal{A}$ defined on

$$\mathcal{D}(\mathcal{A}) = \{(\eta, \varphi) \in \mathbb{R}^n \times L_2 / \varphi \in W^{1,2},\ \varphi(0) = \eta\} \tag{1.3}$$

by

$$\mathcal{A}(\eta, \varphi) = (L\varphi, \dot{\varphi}) \tag{1.4}$$

generated a C_0-semigroup $\mathcal{S}(t)$ on $\mathbb{R}^n \times L_2$. Moreover, for $(\eta, \varphi) \in \mathbb{R}^n \times L_2$,

$$\mathcal{S}(t)(\eta, \varphi) = (x(t), x_t(\cdot)) \tag{1.5}$$

where $x(\cdot)$ is the unique absolutely continuous solution to the homogeneous form (i.e., $f(t) \equiv 0$) of the RFDE (1.1)–(1.2). In 1977, Vinter [30] proved the same results under the milder assumption that L be "C-bounded" and conjectured that the result remained valid for any $L \in \mathcal{B}(W^{1,2}, \mathbb{R}^n)$. In 1978, Delfour [10] proved Vinter's conjecture and in the same paper Delfour stated that the converse was also true: if $\mathcal{A}$ defined by (1.3)–(1.4) generates a C_0-semigroup on $\mathbb{R}^n \times L_2$, then it was necessary that $L \in \mathcal{B}(W^{1,2}, \mathbb{R}^n)$.

Section 2 is devoted to the extension of the aforementioned results to a general class of functional differential equations of the form

$$\frac{d}{dt} Dx_t = Lx_t + f(t). \tag{1.6}$$

where L and D are $\mathbb{R}^n$-valued linear operators. Proofs are given for the results announced by the present authors in the note [8]. In §3 we establish an equivalence between NFDEs of the form (1.6) and an abstract control system $\dot{z} = \mathcal{A}z + \mathcal{T}(t)$ in $\mathbb{R}^n \times L_p$. Finally, §4 is concerned with a class of singular integral equations not covered by the previous theory.

2. Semigroups on product spaces. Let L and D be linear $\mathbb{R}^n$-valued operators with domains $\mathcal{D}(L)$ and $\mathcal{D}(D)$ subspaces of the Lebesgue measurable $\mathbb{R}^n$-valued functions on $[-r,0]$. It is assumed that $W^{1,p}\subseteq\mathcal{D}(L)\cap\mathcal{D}(D)$. However, note that at this point we make no continuity assumptions on L or D. Define the operator $\mathcal{A}$ with domain

$$\mathcal{D}(\mathcal{A})=\{(\eta,\varphi)\in\mathbb{R}^n\times L_p/\varphi\in W^{1,p},\ D\varphi=\eta\},\tag{2.1}$$

by

$$\mathcal{A}(\eta,\varphi)=(L\varphi,\dot{\varphi}).\tag{2.2}$$

Observe that if D is defined on $W^{1,p}$ by $D\varphi=\varphi(0)$, then the operator $\mathcal{A}$ reduces to the operator (defined by (1.3)–(1.4)) studied by Borisovič and Turbabin, Vinter and Delfour. Theorems 2.1 and 2.2 give necessary conditions imposed on the operators L and D if $\mathcal{A}$ is the infinitesimal generator of a C_0-semigroup on $\mathbb{R}^n\times L_p$. Theorems 2.3 and 2.4 give sufficient conditions on L and D that imply the operator $\mathcal{A}$ generates a C_0-semigroup and C_0-group on $\mathbb{R}^n\times L_p$. The majority of this section is devoted to the proof of Theorem 2.3. For the sake of completeness, we give outlines of the proofs of Theorems 2.1 and 2.2. (Detailed proofs of these theorems may be found in [8] for the case $p=2$ and only slight modifications are needed for general p.)

For $\lambda\in\mathbb{C}$ and $\psi\in L_p$, define the operator $M_\lambda: L_p\to L_p$ by

$$[M_\lambda\psi](t)=\int_0^t e^{\lambda(t-u)}\psi(u)\,du.\tag{2.3}$$

The following lemma is useful and quite easy to establish.

LEMMA 2.1. *The operator defined by* (2.3) *has the following properties*:

i) $\mathcal{R}(M_\lambda)=\tilde{W}^{1,p}\equiv\{\varphi\in W^{1,p}/\varphi(0)=0\}$,

ii) $M_\lambda\in\mathcal{B}(L_p,\tilde{W}^{1,p})$ *and*

iii) M_λ^{-1} *exists and* $M_\lambda^{-1}\in\mathcal{B}(\tilde{W}^{1,p},L_p)$.

THEOREM 2.1. *If* $\mathcal{A}$ *defined by* (2.1)–(2.2) *is the infinitesimal generator of a* C_0*-semigroup on* $\mathbb{R}^n\times L_p$, *then both* L *and* D *belong to* $\mathcal{B}(W^{1,p},\mathbb{R}^n)$.

Outline of proof. As a consequence of the Hille–Yosida theorem [9], [27] we have that $\mathcal{A}$ is densely defined, closed and there exists a real ω such that $\lambda\in\rho(\mathcal{A})$ for all $\lambda>\omega$. In particular, for each $(\xi,\psi)\in\mathbb{R}^n\times L_p$ the equation

$$(\mathcal{A}-\lambda I)(\eta,\varphi)=(\xi,\psi)\tag{2.4}$$

has a unique solution $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$ and the solution depends continuously on (ξ,ψ). Consequently, it follows that there is a constant vector c such that $\varphi(s)=e^{\lambda s}c+[M_\lambda\psi](s)$ and c satisfies

$$\Delta(\lambda)c=-\xi-\lambda D[M_\lambda\psi]+L[M_\lambda\psi]\tag{2.5}$$

where $\Delta(\lambda)=\lambda D[e^{\lambda\cdot}I]-L[e^{\lambda\cdot}I]$. The continuous dependence of φ on (ξ,ψ) together with (2.5) imply that $(\lambda D-L)M_\lambda$ belongs to $\mathcal{B}(L_p,\mathbb{R}^n)$. Therefore, $(\lambda D-L)=(\lambda D-L)M_\lambda M_\lambda^{-1}$ belongs to $\mathcal{B}(\tilde{W}^{1,p},\mathbb{R}^n)$ and since $W^{1,p}\cong\mathbb{R}^n\times\tilde{W}^{1,p}$ we have that $(\lambda D-L)\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ for all $\lambda>\omega$. However, D and L both belong to $\mathcal{B}(W^{1,p},\mathbb{R}^n)$ if and only if $(\lambda D-L)\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ for all $\lambda>\omega$ and the proof is complete. □

If $\mathcal{A}$ generates a C_0-semigroup on $\mathbb{R}^n\times L_p$, we can apply Theorem 2.1 together with standard representation theorems [1] to obtain $n\times n$ matrix-valued functions A,B,F and G whose column vectors belong to L_q $(1/p+1/q=1)$, such that if $\varphi\in W^{1,p}$ then

$$L\varphi=\int_{-r}^0\{F(s)\varphi(s)+G(s)\dot{\varphi}(s)\}\,ds,\tag{2.6}$$

and

$$D\varphi=\int_{-r}^{0}\{A(s)\varphi(s)+B(s)\dot{\varphi}(s)\}\,ds. \tag{2.7}$$

In light of Theorem 2.1, we assume that D and L belong to $\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and the representations (2.6)–(2.7) hold.

THEOREM 2.2. *Assume that $\mathcal{A}$ is the infinitesimal generator of a C_0-semigroup $\{\mathcal{S}(t)\}_{t\geq 0}$ and let D and L have the representation* (2.6)–(2.7). *If $(\eta,\varphi)\in\mathbb{R}^n\times L_p$, then there exist unique functions $y:[0,+\infty)\to\mathbb{R}^n$, $x:[-r,+\infty)\to\mathbb{R}^n$ such that $y(\cdot)$ is continuous, $x_t(\cdot)\in L_p$ for all $t\geq 0$, $x_0(s)=\varphi(s)$ a.e. on $[-r,0]$ and*

$$\mathcal{S}(t)(\eta,\varphi)=(y(t),x_t), \tag{2.8}$$

where

$$y(t)=\eta+\int_{-r}^{0}G(u)\{x(t+u)-x(u)\}\,du+\int_{0}^{t}\int_{-r}^{0}F(u)x(s+u)\,du\,ds. \tag{2.9}$$

Moreover, if $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$, then for each $t\geq 0$, $x_t\in W^{1,p}$, Dx_t is continuously differentiable

$$\mathcal{S}(t)(\eta,\varphi)=(Dx_t,x_t), \tag{2.10}$$

and $x(\cdot)$ is the unique $W^{1,p}$ solution to the NFDE

$$\frac{d}{dt}Dx_t=Lx_t, \tag{2.11}$$

with initial data

$$x_0(s)=\varphi(s),\qquad -r\leq s\leq 0. \tag{2.12}$$

Outline of proof. First assume that $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$. It follows that $z(t)=\mathcal{S}(t)(\eta,\varphi)=(z_1(t),z_2(t,\cdot))$ is the unique continuously differentiable solution to the Cauchy problem

$$\dot{z}(t)=\mathcal{A}z(t),\qquad z(0)=(\eta,\varphi). \tag{2.13}$$

The Cauchy problem (2.13) is equivalent to the system

$$\frac{d}{dt}z_1(t)=Lz_2(t,\cdot),\qquad z_1(0)=\eta, \tag{2.14}$$

$$\frac{\partial}{\partial t}z_2(t,s)=\frac{\partial}{\partial s}z_2(t,s),\qquad z_2(0,s)=\varphi(s), \tag{2.15}$$

where $t\geq 0$, $-r\leq s\leq 0$ and $z_2(t,\cdot)\in W^{1,p}$. Equation (2.15) implies that there exists a function $x:[-r,+\infty)\to\mathbb{R}^n$ such that x is absolutely continuous on compact subintervals of $[-r,+\infty)$, $x_0=\varphi$ and $z_2(t,s)=x(t+s)$. Since $z(t)=(z_2(t),z_2(t,\cdot))=(z_1(t),x_t(\cdot))\in\mathcal{D}(\mathcal{A})$, it follows that $x_t\in W^{1,p}$ and $Dx_t=z_1(t)$ is continuously differentiable. In particular, x is the unique $W^{1,p}$ solution of the initial value problem (2.11)–(2.12), or equivalently,

$$\begin{aligned}Dx_t&=\eta+\int_0^t\int_{-r}^0F(u)x(s+u)\,du\,ds+\int_0^t\int_{-r}^0G(u)\dot{x}(s+u)\,du\,ds\\&=\eta+\int_0^t\int_{-r}^0F(u)x(s+u)\,du\,ds+\int_{-r}^0G(u)\{x(t+u)-x(u)\}\,du\\&=y(t).\end{aligned}$$

For the general case where $(\eta,\varphi)\in\mathbb{R}^n\times L_p$, select a sequence $(\eta^N,\varphi^N)\in\mathcal{D}(\mathcal{A})$ such that $(\eta^N,\varphi^N)\to(\eta,\varphi)$ and let x^N denote the unique solution to the NFDE (2.10) satisfying $x_0^N=\varphi^N$. Strong continuity of $\mathcal{S}(t)$ implies that $\lim_{N\to+\infty}(Dx_t^N,x_t^N)=\mathcal{S}(t)(\eta^N,\varphi^N)=(y_1(t),y_2(t,\cdot))$, where $y_1(t)$ is continuous and $y_2(t,\cdot)\in L_p$. It follows that $\{x^N\}$ is a Cauchy sequence in $L_p(-r,t)$ for all $t\geq 0$ and hence there is a unique function $x:[-r,+\infty)\to\mathbb{R}^n$ such that $x_t\in L_p$ and $\int_{-r}^t\|x^N(s)-x(s)\|^p\,ds\to 0$ for each $t\geq 0$. Consequently, $x_t^N\to x_t$ and $x_t=y_2(t,\cdot)$ in L_p. The convergence of x_t^N to x_t and φ^N to φ can then be employed to show that $y_1(t)=y(t)$ where $y(t)$ is defined by (2.9). □

Remark 2.1. It follows directly that if $\mathcal{A}$ generates a C_0-semigroup on $\mathbb{R}^n\times L_p$, then D *cannot* be bounded as an operator on L_p (i.e., $D\notin\mathcal{B}(L_p,\mathbb{R}^n)$). In fact, if D did belong to $\mathcal{B}(L_p,\mathbb{R}^n)$ then by the density of $\mathcal{D}(\mathcal{A})$ the representation (2.10) would hold for all $(\eta,\varphi)\in\mathbb{R}^n\times L_p$. Strong continuity of $\mathcal{S}(t)$ at $t=0$ would imply that for each $(\eta,\varphi)\in\mathbb{R}^n\times L_p$, $\lim_{t\to 0^+}(Dx_t,x_t)=(\eta,\varphi)$, which leads to the contradiction

$$\eta=\lim_{t\to 0^+}Dx_t=D\varphi$$

for all $(\eta,\varphi)\in\mathbb{R}^n\times L_p$. This observation leads to the fact that the matrix valued function B in the representation (2.7) can not be zero a.e. on $[-r,0]$.

Our attention will now be focused upon conditions on L and D which are sufficient for $\mathcal{A}$ to generate a C_0-semigroup on $\mathbb{R}^n\times L_p$.

Prior to the statement of the principal result (Theorem 2.3 below) we recall the following representations and definitions. If $D\in\mathcal{B}(C,\mathbb{R}^n)$, then there is a $n\times n$ matrix-valued function $\mu(\cdot)$ whose entries are of bounded variation on $[-r,0]$ and such that if $\varphi\in C$, then

$$D\varphi=\int_{-r}^0 d\mu(s)\varphi(s).$$

We shall assume that $\mu(\cdot)$ is normalized to be right continuous on $(-r,0)$ with $\mu(-r)=0$ and extend $\mu(\cdot)$ over $\mathbb{R}$ by $\mu(s)=\mu(0)$ for $s\geq 0$ and $\mu(s)=0$ for $s\leq -r$. The operator D is said to be atomic at $s\in[-r,0]$ if the jump at s, $J(\mu,s)\equiv\mu(s)-\mu(s^-)$ is nonsingular. In the case where D is atomic at zero, we may assume without loss of generality that

$$D\varphi=\varphi(0)+\int_{-r}^0 d\mu(s)\varphi(s) \tag{2.16}$$

and

$$\lim_{\varepsilon\to 0}\operatorname{Var}_{[-\varepsilon,0]}(\mu)=0, \tag{2.17}$$

where $\operatorname{Var}_{[a,b]}(\mu)$ denotes the total variation of $\mu(\cdot)$ on $[a,b]$.

THEOREM 2.3. *If* $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ *and* $D\in\mathcal{B}(C,\mathbb{R}^n)$ *has an atom at* $s=0$, *then* $\mathcal{A}$ *defined by* (2.1)–(2.2) *generates a* C_0*-semigroup on* $\mathbb{R}^n\times L_p$.

Example 2.1. In order to illustrate the types of neutral equations covered by Theorem 2.3, we present a couple of specific examples. In particular, let

$$L\varphi=\sum_{j=0}^m A_j\varphi(-r_j)+\int_{-r}^0 A(s)\varphi(s)\,ds$$

and

$$D\varphi=\varphi(0)-\sum_{j=1}^m B_j\varphi(-r_j)+\int_{-r}^0 B(s)\varphi(s)\,ds,$$

where $0=r_0<r_1<\cdots<r_m=r$ and $A(\cdot)$, $B(\cdot)$ are in $L_2((-r,0);\mathbb{R}^{n\times n})$. Clearly, $L\in\mathcal{B}(W^{1,2},\mathbb{R}^n)$, $D\in\mathcal{B}(C,\mathbb{R}^n)$ and D is atomic at $s=0$ so that Theorem 2.3 applies. This special case was considered by Kappel in [21]. Using a norm on $\mathbb{R}^n\times L_2$ equivalent to the usual norm, Kappel was able to show in this case that $(\mathcal{A}-\omega I)$ is maximal dissipative and hence the generator of a C_0-semigroup on $\mathbb{R}^n\times L_2$.

As another example, let $r=1$, $n=1$ and define the operators

$$L\varphi=\sum_{j=0}^{\infty}a_j\varphi(-r_j)+\int_{-1}^{0}|s|^{-1/4}\dot{\varphi}(s)\,ds$$

and

$$D\varphi=\varphi(0)-\sum_{j=1}^{\infty}b_j\varphi(-r_j)+\int_{-1}^{0}b(s)\varphi(s)\,ds,$$

where $0=r_0<r_j\le 1$, $j=1,2,3,\cdots$, $\Sigma_{j=1}^{\infty}|b_j|^2<+\infty$, $\Sigma_{j=0}^{\infty}|a_j|^2<+\infty$ and $b(\cdot)\in L_2$. Observe that the operators L and D define a neutral functional differential equation with an infinite number of delays. It is not difficult to show that the operator L does not belong to $\mathcal{B}(C,R^1)$. Therefore, the theory in [15] does not apply. However, Theorem 2.3 is applicable since $L\in\mathcal{B}(W^{1,2},R^1)$, $D\in\mathcal{B}(C,R^1)$ and D is atomic at $s=0$.

Theorem 2.3 is an immediate consequence of the technical Lemmas (2.1)–(2.8) given below. The proof of Theorem 2.3 will be delayed until these lemmas are established. The following result may be found in the paper [5] by Brown and Krall.

LEMMA 2.1. *Let $B_1, B_2,\cdots,B_n$ be a finite collection of linear functionals on a normed linear space X. Then $\cap_{i=1}^{n}\mathcal{N}(B_i)$ is dense in X if and only if every nonzero linear combination $\Sigma_{i=1}^{n}\lambda_i B_i$, $\lambda_i\in\mathbb{R}$ is unbounded.*

LEMMA 2.2. *If $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and $D\in\mathcal{B}(C,\mathbb{R}^n)$ is atomic at $s=0$, then $\mathcal{D}(\mathcal{A})$ defined by* (2.1) *is dense in $\mathbb{R}^n\times L_p$.*

Proof. Define the operator $B:\mathbb{R}^n\times W^{1,p}\to\mathbb{R}^n$ by $B(\eta,\varphi)=\eta-D\varphi$ and let $B_i(\eta,\varphi)$ denote the ith coordinate of $B(\eta,\varphi)$, $i=1,2,\cdots,n$. Clearly, $\mathcal{D}(\mathcal{A})=\cap_{i=1}^{n}\mathcal{N}(B_i)$ and the lemma will be proved if we can show that each nonzero linear combination

$$\Lambda=\sum_{i=1}^{n}\lambda_i B_i$$

is unbounded on $\mathbb{R}^n\times L_p$. Without loss of generality we may take $\lambda_1=1$ and assume that D has the representations (2.16)–(2.17).

For $0<\varepsilon<r$ define $\gamma(\varepsilon)=\int_{-\varepsilon}^{0}|d\mu(s)|=\mathrm{Var}_{[-\varepsilon,0]}(\mu)$ and note that $\lim_{\varepsilon\to 0}\gamma(\varepsilon)=0$. If $\gamma(\varepsilon)=0$ for any $\varepsilon>0$, the discontinuity of Λ is immediate because point evaluation at 0 is undefined. If $\gamma(\varepsilon)>0$ on $(0,r)$, then define $\{\varphi^\varepsilon(\cdot)\}_{\varepsilon>0}$ to be any family of smooth real-valued functions for which $\varphi^\varepsilon(s)\equiv 0$ on $[-r,-\varepsilon]$, $\varphi^\varepsilon(\cdot)$ is nondecreasing on $[-r,0]$, $\varphi^\varepsilon(0)=1/\gamma(\varepsilon)$ and $\varphi^\varepsilon(\cdot)\to 0$ in $L_p([-r,0];\mathbb{R})$ as $\varepsilon\to 0$. Define $\psi^\varepsilon:[-r,0]\to\mathbb{R}^n$ by $\psi^\varepsilon(s)=e_1\varphi^\varepsilon(s)$ where $\{e_i\}_{i=1}^{n}$ is the standard basis for $\mathbb{R}^n$. It follows that $\psi^\varepsilon(\cdot)\to 0$ in L_p and yet

$$\begin{aligned}|\Lambda(0,\psi^\varepsilon)|&=\left|\varphi^\varepsilon(0)+\sum_{i=1}^{n}\lambda_i\int_{-\varepsilon}^{0}\langle e_i,d\mu(s)\psi^\varepsilon(s)\rangle\,ds\right|\\&\ge\varphi^\varepsilon(0)-\sum_{i=1}^{n}|\lambda_i|\int_{-\varepsilon}^{0}|d\mu(s)|\,\|\psi^\varepsilon\|_C\\&\ge 1/\gamma(\varepsilon)-\left(\sum_{i=1}^{n}|\lambda_i|\right)\gamma(\varepsilon)\cdot\gamma^{-1}(\varepsilon)\\&\ge 1/\gamma(\varepsilon)-\text{constant}.\end{aligned}$$

The conclusion follows immediately. □

LEMMA 2.3. *Let L and D belong to $\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and $\mathcal{A}$ be given by* (2.1)–(2.2). *Then $\lambda\in\rho(\mathcal{A})$ if and only if $\Delta(\lambda)\equiv\lambda D(e^{\lambda\cdot}I)-L(e^{\lambda\cdot}I)$ is nonsingular. If, in addition, $D\in\mathcal{B}(C,\mathbb{R}^n)$ is atomic at $s=0$, then $\lambda\in\rho(\mathcal{A})$ for all sufficiently large real λ.*

Proof. A real number λ belongs to $\rho(\mathcal{A})$ if and only if the equation

$$(\mathcal{A}-\lambda I)(\eta,\varphi)=(\xi,\psi) \tag{2.18}$$

has a unique solution $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$ for each $(\xi,\psi)\in\mathbb{R}^n\times L_p$ and that solution depends continuously on (ξ,ψ). The second coordinate of (2.18) is equivalent to

$$\varphi(s)=e^{\lambda s}c+[M_\lambda\psi](s),\qquad -r\le s\le 0,$$

where the constant vector c is determined by the first coordinate of (2.18) and the requirement that $D\varphi=\eta$, i.e.,

$$L(e^{\lambda\cdot}I)c-\lambda\eta=\xi-L(M_\lambda\psi)$$

and

$$D(e^{\lambda\cdot}I)c-\eta=-D(M_\lambda\psi).$$

This system is uniquely solvable for c and η if and only if $\Delta(\lambda)$ is nonsingular. In the case that $\det\Delta(\lambda)\neq 0$, the continuous dependence of (η,φ) on (ξ,ψ) follows from the fact that LM_λ and DM_λ belong to $\mathcal{B}(L_p;\mathbb{R}^n)$ and $M_\lambda\in\mathcal{B}(L_p,L_p)$.

To prove the second assertion we first note (using the representation (2.16)–(2.17)) that $\lim_{\lambda\to+\infty}D(e^{\lambda\cdot}I)=I$. Since $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ it follows that $\|L(e^{\lambda\cdot}I)\|\le\|L\|\,(1+\lambda)(p\lambda)^{-1/p}$, which implies that $\|\lambda^{-1}L(e^{\lambda\cdot}I)\|\to 0$ as $\lambda\to+\infty$. Consequently, $\lambda^{-1}\Delta(\lambda)=D(e^{\lambda\cdot}I)+\lambda^{-1}L(e^{\lambda\cdot}I)\to I$ as $\lambda\to+\infty$ and for sufficiently large λ, the matrix $\lambda^{-1}\Delta(\lambda)$ (hence $\Delta(\lambda)$) is nonsingular. □

LEMMA 2.4. *Let μ be as in* (2.16). (1) *If $x\in L_p(\mathbb{R})$, then*

$$[\mu*x](t)=\int_{-\infty}^{+\infty}d\mu(s)x(t+s)$$

is finite a.e. on $\mathbb{R}$, $\mu*x\in L_p(\mathbb{R})$ and

$$\|\mu*x\|_{L_p(\mathbb{R})}\le\operatorname*{Var}_{\mathbb{R}}(\mu)\cdot\|x\|_{L_p(\mathbb{R})}.$$

(2) *If $x\in W^{1,p}(-r,a)$ for some $a>0$, then $\mu*x\in W^{1,p}(0,a)$, $\frac{d}{dt}[\mu*x]=\mu*\dot{x}$ a.e. on $[0,a]$ and there is a constant K (independent of a,r,μ, and x) such that*

$$\|\mu*x\|_{W^{1,p}(0,a)}\le K\operatorname*{Var}_{\mathbb{R}}(\mu)\|x\|_{W^{1,p}(-r,a)}.$$

Proof. Part (1) is a special case of Hewitt and Ross [18, Thm. 20.12]; therefore we concentrate on the proof of (2). There exists an operator $E\in\mathcal{B}(W^{1,p}(-r,a),W^{1,p}(\mathbb{R}))$ whose norm K is independent of r and a such that $[Ex](s)=x(s)$ for each $s\in[-r,a]$ (see [1]). We write $\hat{x}=Ex$ and note that without loss of generality $\hat{x}$ may be taken to have compact support.

Let $\{T(t)\}_{t\ge 0}$ be the C_0-semigroup of left translations on $L_p(\mathbb{R})$, i.e., $[T(t)z](s)=z(t+s)$ for $s\in\mathbb{R}$, $z\in L_p(\mathbb{R})$. The infinitesimal generator Ω for this semigroup is the operator defined on its domain $\mathcal{D}(\Omega)=\{z\in L_p(\mathbb{R})/\dot{z}\in L_p(\mathbb{R})\}$ by $[\Omega z](t)=\dot{z}(t)$ (see [9]).

Let $x\in W^{1,p}(-r,a)$ and define $y=\mu*\hat{x}$ and $z=\mu*\dot{\hat{x}}$. By part (1), both y and z belong to $L_p(\mathbb{R})$ and $\|z-1/h\{y_h-y\}\|_{L_p(\mathbb{R})}\le\operatorname{Var}_{\mathbb{R}}(\mu)\|\dot{\hat{x}}-1/h\{\hat{x}_h-\hat{x}\}\|_{L_p(\mathbb{R})}$. Since $\hat{x}\in\mathcal{D}(\Omega)$, this inequality implies that $1/h\{y_h-y\}=1/h\{T(h)-I\}y\to z$ and hence $y\in\mathcal{D}(\Omega)$ with $\Omega y=z$. If $t\in[0,a]$, then

$$y(t)=\int_{-\infty}^{+\infty}d\mu(s)\hat{x}(t+s)=\int_{-r}^{0}d\mu(s)\hat{x}(t+s)=\int_{-r}^{0}d\mu(s)x(t+s)$$

and

$$\dot{y}(t)=[\Omega y](t)=z(t)=\int_{-\infty}^{+\infty} d\mu(s)\dot{x}(t+s)=\int_{-r}^{0} d\mu(s)\dot{x}(t+s).$$

Finally, by part (1)

$$\begin{aligned}\|\mu * x\|_{W^{1,p}(0,a)} &\le \|y\|_{W^{1,p}(\mathbb{R})}=\|y\|_{L_p(\mathbb{R})}+\|z\|_{L_p(\mathbb{R})}\\ &\le \operatorname*{Var}_{\mathbb{R}}(\mu)\left[\|\hat{x}\|_{L_p(\mathbb{R})}+\|\dot{\hat{x}}\|_{L_p(\mathbb{R})}\right]\\ &\le K \operatorname*{Var}_{[-r,0]}(\mu)\|x\|_{W^{1,p}(-r,a)}.\end{aligned}$$

□

LEMMA 2.5. *Let $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ have the representation* (2.6) *and $D\in\mathcal{B}(C,\mathbb{R}^n)$ be atomic at $s=0$ with the representation* (2.16)–(2.17). (1) *If $-r<-\varepsilon<0<a$ and $z\in L_p(-r,a)$ satisfies $z(t)=0$ a.e. on $[-r,-\varepsilon]$, then*

$$\left\|\int_{-r}^{0} d\mu(s)z(\cdot+s)-\int_{-r}^{0} G(s)z(\cdot+s)\,ds\right\|_{L_p(0,a)} \le \left[\operatorname*{Var}_{[-\varepsilon,0]}(\mu)+\int_{-\varepsilon}^{0}\|G(s)\|\,ds\right]\|z\|_{L_p(-r,a)}.$$

(2) *If $-r<-\varepsilon<0<a$ and $z\in W^{1,p}(-r,a)$ satisfies $z(t)=0$ on $[-r,-\varepsilon]$, then*

$$\left\|\int_{-r}^{0} d\mu(s)z(\cdot+s)-\int_{-r}^{0} G(s)z(\cdot+s)\right\|_{W^{1,p}(0,a)} \le K\left[\operatorname*{Var}_{[-\varepsilon,0]}(\mu)+\int_{-\varepsilon}^{0}\|G(s)\|\,ds\right]\|z\|_{W^{1,p}(-r,a)},$$

where K is the constant in Lemma 2.4.

Proof. Define $\tilde{\mu}:\mathbb{R}\to\mathbb{R}^{n\times n}$ by

$$\tilde{\mu}(s)=\begin{cases}\mu(-\varepsilon)+\int_{-\varepsilon}^{0} G(u)\,du, & s\le-\varepsilon,\\ \mu(s)+\int_{s}^{0} G(u)\,du, & -\varepsilon\le s<0,\\ \mu(0^-), & 0\le s.\end{cases}\tag{2.19}$$

If $z(s)=0$ for $-r\le s\le-\varepsilon$, then for $0\le t\le a$

$$\int_{-r}^{0} d\tilde{\mu}(s)z(t+s)=\int_{-r}^{0} d\mu(s)z(t+s)-\int_{-r}^{0} G(s)z(t+s)\,ds.$$

The conclusions (1) and (2) now follow from the previous lemma and the observation that

$$\operatorname*{Var}_{[-r,0]}(\tilde{\mu})\le\operatorname*{Var}_{[-\varepsilon,0]}(\mu)+\int_{-\varepsilon}^{0}\|G(s)\|\,ds.$$

The next two results establish the basic existence and continuous dependence for solutions and generalized solutions to the NFDE $\frac{d}{dt}Dx_t=Lx_t+f(t)$ with initial data $(\eta,\varphi)\in\mathbb{R}^n\times L_p$.

LEMMA 2.6. *Let $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ have the representation* (2.6) *and $D\in\mathcal{B}(C,\mathbb{R}^n)$ be atomic at $s=0$ with the representation* (2.16)–(2.17). *If $(\eta,\varphi)\in\mathbb{R}^n\times L_p$ and $f\in L_p^{\mathrm{loc}}(0,+\infty)$, then there exists a unique (in the a.e. sense) function $x:[-r,+\infty)\to\mathbb{R}^n$ such that $x_0(s)=\varphi(s)$ a.e. on $[-r,0]$, $x_t\in L_p$ for $t\ge0$ and*

$$\begin{aligned}x(t)+\int_{-r}^{0} d\mu(s)x(t+s)=\eta&+\int_{-r}^{0} G(s)\{x(t+s)-x(s)\}\,ds\\ &+\int_0^t\int_{-r}^{0} F(s)x(u+s)\,ds\,du+\int_0^t f(u)\,du\end{aligned}\tag{2.20}$$

for a.e. $t\ge0$.

Moreover, if $\varphi \in W^{1,p}$ *and* $\eta = D\varphi$, *then* $x_t \in W^{1,p}$, $Dx_t \in W^{1,p}$ *for* $t \geq 0$ *and*

$$\frac{d}{dt} Dx_t = Lx_t + f(t), \tag{2.21}$$

for a.e. $t \geq 0$. *If in addition* f *is continuous, then* Dx_t *is continuously differentiable.*

Proof. The proof is based on a fixed point argument. For $0 < a < r$, let $\varepsilon = a$ and $\tilde{\mu}$ be the measure defined by (2.19). Let $T = T(\eta, \varphi, f)$ be the operator defined on $L_p(-r, a)$ by

$$[Tx](t) = \begin{cases} \varphi(t), & -r \leq t < 0, \\ \eta + \int_0^t f(u)\,du - \int_{-r}^0 G(u)\varphi(u)\,du \\ \quad + \int_0^t \int_{-r}^0 F(s)x(u+s)\,ds\,du \\ \quad + \int_{-r}^0 G(s)x(t+s)\,ds - \int_{-r}^0 d\mu(s)x(t+s), & 0 \leq t \leq a. \end{cases} \tag{2.22}$$

Lemma 2.4, part (1) implies that $T: L_p(-r, a) \to L_p(-r, a)$ and if $x, y \in L_p(-r, a)$, then Lemma 2.4, part (1), and the convolution theorem [19] imply that

$$\|Tx - Ty\|_{L_p(-r,a)} \leq M_1(r)\|x - y\|_{L_p(-r,a)}, \tag{2.23}$$

where $M_1(\tau)$ is the function defined by

$$M_1(\tau) = \left[\operatorname*{Var}_{[-\tau,0]} (\mu) + \int_{-\tau}^0 \|G(s)\|\,ds + \tau^{1/p} \int_{-\tau}^0 \|F(s)\|\,ds \right].$$

Let $U = U(\eta, \varphi, f) = T^2$. We shall show that U is a contraction on $L_p(-r, a)$ (for sufficiently small a) and make use of the observation that the resulting unique fixed point of U is also a unique fixed point of T (see Lemma 5.4.3 in [23]). Let $x, y \in L_p(-r, a)$ and define $\hat{x} = Tx$, $\hat{y} = Ty$. Since $[Tx](t) = [Ty](t) = \varphi(t)$ for $-r \leq t \leq 0$, the function $\hat{z} = \hat{x} - \hat{y}$ is zero on $[-r, 0]$. By the definition of $\tilde{\mu}$ we have that

$$[T\hat{x} - T\hat{y}](t) = \begin{cases} 0, & -r \leq t \leq 0, \\ -\int_{-r}^0 d\tilde{\mu}(s)\hat{z}(t+s) + \int_0^t \int_{-r}^0 F(s)\hat{z}(u+s)\,ds\,du, & 0 \leq t \leq a. \end{cases} \tag{2.24}$$

The function $\hat{z}$ satisfies the assumption in Lemma 2.5, part (1), which implies that

$$\|T\hat{x} - T\hat{y}\|_{L_p(-r,a)} \leq M_1(a)\|\hat{x} - \hat{y}\|_{L_p(-r,a)}. \tag{2.25}$$

Using the fact that $Ux - Uy = T\hat{x} - T\hat{y}$ and combining inequalities (2.23) and (2.25), we have that

$$\|Ux - Uy\|_{L_p(-r,a)} \leq M_1(a)M_1(r)\|x - y\|_{L_p(-r,a)}.$$

By (2.17), $M_1(a) \to 0$ as $a \to 0$ and, therefore, U will be a contraction on $L_p(-r, a)$ with Lipschitz constant $\alpha = M_1(a)M_1(r)$ independent of (η, φ, f). The unique fixed point $x: [-r, a] \to \mathbb{R}^n$ is also a fixed point of T and defines a unique solution to (2.20). The independence of a on (η, φ, f) allows one to establish the existence of solutions on $[a, 2a], [2a, 2a], \cdots$.

If $\varphi \in W^{1,p}$, then we define $T = T(D\varphi, \varphi, f)$ by (2.22). Lemma 2.4, part (2), and the convolution theorem [19] imply that $T: W^{1,p}(-r, a) \to W^{1,p}(-r, a)$. If $x, y \in W^{1,p}(-r, a)$, then

$$\|Tx - Ty\|_{W^{1,p}(-r,a)} \leq M_2(r)\|x - y\|_{W^{1,p}(-r,a)}, \tag{2.26}$$

where $M_2(\tau)$ is the function

$$M_2(\tau)=\left\{K\left[\operatorname*{Var}_{[-\tau,0]}(\mu)+\int_{-\tau}^{0}\|G(s)\|\,ds\right]+\left[\int_{-\tau}^{0}\|F(s)\|\,ds\right](1+\tau^{1/p})\right\},$$

and K is the constant from Lemma 2.4. Let $U=U(D\varphi,\varphi,f)=T^2$ and for $x,y\in W^{1,p}(-r,a)$ define $\hat{x}=Tx$, $\hat{y}=Ty$. As before, $\hat{z}=\hat{x}-\hat{y}$ satisfies the assumption in Lemma 2.5, part (2). Consequently, it follows that

$$\|T\hat{x}-T\hat{y}\|_{W^{1,p}(-r,a)}\le M_2(a)\|\hat{x}-\hat{y}\|_{W^{1,p}(-r,a)}, \tag{2.27}$$

and (2.26)–(2.27) combine to yield

$$\|Ux-Uy\|_{W^{1,p}(-r,a)}\le M_2(a)M_2(r)\|x-y\|_{W^{1,p}(-r,a)}.$$

The contraction mapping principle can again be applied to obtain $x\in W^{1,p}(-r,a)$ which solves (2.20) and x can be extended to $[-r,+\infty)$ such that $x_t\in W^{1,p}$ for all $t\ge 0$. By Lemma 2.4, $Dx_t\in W^{1,p}(0,a)$ for all $a>0$ and differentiation of (2.20) yields $\frac{d}{dt}Dx_t=Lx_t+f(t)$ for a.e. $t\ge 0$.

If in addition f is continuous, then $Lx_t+f(t)$ is continuous in t (by the convolution theorem). Hence $\frac{d}{dt}Dx_t$ is continuous, which completes the proof. □

Let Γ denote the product space $\Gamma=\mathbb{R}^n\times L_p\times L_p(0,a)$ and observe that for $\gamma=(\eta,\varphi,f)$ the mappings $U=U(\gamma)=T^2(\gamma)$ defines a uniform contraction on $L_p(-r,a)$ (see [14, p. 6]). Moreover, if $\gamma=(\eta,\varphi,f)$ and $\lambda=(\xi,\psi,g)$ belong to Γ, then for $x\in L_p(-r,a)$

$$[T(\gamma)x-T(\lambda)x](t)=\begin{cases}\varphi(t)-\psi(t), & -r\le t<0,\\ (\eta-\xi)+\int_0^t\{f(s)-g(s)\}\,ds & \\ \qquad -\int_{-r}^{0}G(s)\{\varphi(s)-\psi(s)\}\,ds, & 0\le t\le a.\end{cases} \tag{2.28}$$

Consequently, there is a constant Δ (independent of x) such that

$$\|T(\gamma)x-T(\lambda)x\|_{L_p(-r,a)}\le\Delta\|\gamma-\lambda\|_\Gamma \tag{2.29}$$

for all $\gamma,\lambda\in\Gamma$. Using the identity

$$\begin{aligned}U(\gamma)x-U(\lambda)x&=T(\gamma)(T(\gamma)x)-T(\lambda)(T(\lambda)x)\\&=T(\gamma)(T(\gamma)x)-T(\gamma)(T(\lambda)x)+T(\gamma)(T(\lambda)x)-T(\lambda)(T(\lambda)x)\end{aligned}$$

and inequalities (2.23) and (2.29), it follows that

$$\|U(\gamma)x-U(\lambda)x\|_{L_p(-r,a)}\le[M_1(\gamma)\Delta+\Delta]\|\gamma-\lambda\|_\Gamma.$$

Therefore, the family $\{U(\gamma)/\gamma\in\Gamma\}$ is continuous in γ and defines a uniform contraction on $L_p(-r,a)$. An application of [14, Thm. 3.2] yields the following continuous dependence result.

LEMMA 2.7. *If $x(\cdot;\eta,\varphi,f)$ denotes the unique solution of* (2.20) *and $t_1>0$, then the mapping $(\eta,\varphi,f)\to x(\cdot;\eta,\varphi,f)$ is continuous as a function from $\mathbb{R}^n\times L_p\times L_p(0,t_1)$ into $L_p(0,t_1)$.*

LEMMA 2.8. *Let $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and $D\in\mathcal{B}(C,\mathbb{R}^n)$ be atomic at $s=0$. If $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$ and $f\in L_p^{\mathrm{loc}}(0,+\infty)$, then the Cauchy problem*

$$\dot{z}(t)=\mathcal{A}z(t)+(f(t),0), \tag{2.30}$$

$$z(0)=(\eta,\varphi) \tag{2.31}$$

has a unique solution $z(t;\eta,\varphi,f)$. *If* f *is continuous, then* z *is continuously differentiable on* $[0,+\infty)$.

Proof. If $(\eta,\varphi)\in\mathcal{D}(\mathcal{A})$, then by Lemma 2.6 the solution x to (2.20) belongs to $W^{1,p}([0,a];\mathbb{R}^n)$ for all $a>0$ and satisfies the NFDE (2.21). Let $z(t)=(Dx_t,x_t)$ and observe that $z(t)\in\mathcal{D}(\mathcal{A})$ solves the Cauchy problem (2.30)–(2.31). Lemma 2.6 also implies that Dx_t is continuously differentiable if f is continuous. Since $x_t\in W^{1,p}$ for all $t\geq 0$, the mapping $\mathcal{F}:[0,+\infty)\to L_p$ defined by $\mathcal{F}(t)=x_t$ is differentiable, with derivative $\frac{d}{dt}\mathcal{F}(t)=\dot{x}_t$. Moreover, $\|\frac{d}{dt}\mathcal{F}(t_1)-\frac{d}{dt}\mathcal{F}(t_2)\|_{L_p}=\|\dot{x}_{t_1}-\dot{x}_{t_2}\|_{L_p}$, and since translation is a continuous operation on L_p it follows that $\frac{d}{dt}\mathcal{F}(t)$ is continuous. Consequently, if f is continuous then $z(t)=(Dx_t,x_t)$ is continuously differentiable.

Concerning the uniqueness of the solution to (2.30)–(2.31), it suffices to consider the case $f\equiv 0$. If $z(t)=(z_1(t),z_2(t,\cdot))$ satisfies (2.30)–(2.31), then an argument like that given in the proof of Theorem 2.2. shows that $z_2(t,s)=x_t(s)$ where x solves (2.21) and $z_1(t)=Dx_t$. This establishes the uniqueness of z and completes the proof. □

We are now able to prove our sufficiency result.

Proof of Theorem 2.3. The proof makes use of a result due to Phillips [27] which states that an operator $\mathcal{A}$ generates a C_0-semigroup on a Banach space X if and only if (i) $\mathcal{D}(\mathcal{A})$ is dense in X, (ii) the resolvent set $\rho(\mathcal{A})$ is nonempty and (iii) the Cauchy problem

$$\frac{dz(t)}{dt}=\mathcal{A}z(t),\qquad z(0)=z_0\in\mathcal{D}(\mathcal{A})$$

has a unique continuously differentiable solution on $[0,+\infty)$.

If $\mathcal{A}$ is defined by (2.1)–(2.2), then $\mathcal{A}$ is densely defined by Lemma 2.2. Lemma 2.3 implies that $\rho(\mathcal{A})$ is nonempty and the existence of a unique continuously differentiable solution of the Cauchy problem was established in Lemma 2.8. □

Remark. 2.2. If both L and D belong to $\mathcal{B}(C,\mathbb{R}^n)$ with D atomic at $s=0$, then it is well known that the NFDE $\frac{d}{dt}Dx_t=Lx_t$ with continuous initial data $x_0=\varphi\in C$ has a continuously differentiable solution x. The corresponding solution operator $T(t)\colon C\to C$ defined by $T(t)\varphi=x_t$, $t\geq 0$, defines a C_0-semigroup on C (see [15]). By elementary arguments (see [8]) one can show that if $\mathcal{S}(t)$ is the semigroup generated by $\mathcal{A}$, then $\mathcal{S}(t)(D\varphi,\varphi)=(DT(t)\varphi,T(t)\varphi)$ for all $t\geq 0$. Thus, $\{\mathcal{S}(t)\}_{t\geq 0}$ provides an "extension" of the semigroup $\{T(t)\}_{t\geq 0}$ on C.

We conclude this section with a result concerning the solvability of the NFDE for $t\leq 0$.

THEOREM 2.4. *If* $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ *and* $D\in\mathcal{B}(C,\mathbb{R}^n)$ *is atomic at* $s=0$ *and* $s=-r$, *then* $-\mathcal{A}$ *generates a* C_0*-semigroup on* $\mathbb{R}^n\times L_p$. *Consequently,* $\mathcal{A}$ *is the generator of a* C_0*-group on* $\mathbb{R}^n\times L_p$.

Proof. We define the "reflection" operator on L_p by $[R\varphi](s)=\varphi(-s-r)$, for $-r\leq s\leq 0$. Let $Q:\mathbb{R}^n\times L_p\to\mathbb{R}^n\times L_p$ be defined by $Q(\eta,\varphi)=(\eta,R\varphi)$ and note that $Q=Q^{-1}$. Define $L_1\in\mathcal{B}(W^{1,p},R^n)$ and $D_1\in\mathcal{B}(C,\mathbb{R}^n)$ by $L_1=-LR$ and $D_1=DR$, respectively. The operator D_1 is atomic at $s=0$ and Theorem 2.3 may be applied to the operator $\mathcal{A}_1$ defined on

$$\mathcal{D}(\mathcal{A}_1)=\{(\eta,\varphi)\in\mathbb{R}^n\times L_p/\varphi\in W^{1,p},\ D_1\varphi=\eta\}$$

by

$$\mathcal{A}_1(\eta,\varphi)=(L_1\varphi,\dot{\varphi}).$$

In particular, $\mathcal{A}_1$ generates a C_0-semigroup $\{\mathcal{S}_1(t)\}_{t\geq 0}$ on $\mathbb{R}^n\times L_p$.

If we define $\hat{\mathcal{A}}=Q\mathcal{A}_1Q^{-1}$ and $\hat{\mathcal{S}}(t)=Q\mathcal{S}_1(t)Q^{-1}$, then it is easy to show that $\{\hat{\mathcal{S}}(t)\}_{t\geq 0}$ is a C_0-semigroup on $\mathbb{R}^n\times L_p$ with generator $\hat{\mathcal{A}}$. Moreover,

$$\mathcal{D}(\hat{\mathcal{A}})=Q(\mathcal{D}(\mathcal{A}_1))=\mathcal{D}(\mathcal{A})=\mathcal{D}(-\mathcal{A})$$

and $\hat{\mathcal{A}}(\eta,\varphi)=(-L\varphi,-\dot{\varphi})=-\mathcal{A}(\eta,\varphi)$ so that $\hat{\mathcal{A}}=-\mathcal{A}$, which proves that $-\mathcal{A}$ is a generator. The final assertion follows from well-known results (see [27]). □

3. An equivalence theorem. In this section we establish an equivalence between generalized solutions to NFDEs and the mild solutions to the corresponding abstract Cauchy problem. Throughout this section we assume that $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ has the representation (2.6) and $D\in\mathcal{B}(C,\mathbb{R}^n)$ is atomic at $s=0$ with the representation (2.16)–(2.17). The operator $\mathcal{A}$ is defined by (2.1)–(2.2).

We consider the NFDE

$$\frac{d}{dt}Dx_t=Lx_t+f(t),\qquad t>0, \tag{3.1}$$

with initial data

$$Dx_0=\eta,\qquad x_0=\varphi, \tag{3.2}$$

where $(\eta,\varphi)\in\mathbb{R}^n\times L_p$ and $f\in L_p^{\mathrm{loc}}(0,+\infty)$. A solution to (3.1)–(3.2) is a function $x:[-r,+\infty)\to\mathbb{R}^n$ satisfying (i) $x_t\in L_p$ for all $t\geq 0$, (ii) $x(s)=\varphi(s)$ a.e. on $[-r,0]$ and (iii) for a.e. $t\in(0,+\infty)$ x solves the integral equation

$$x(t)+\int_{-r}^{0}d\mu(s)x(t+s)=y(t), \tag{3.3}$$

where

$$\begin{aligned}y(t)=\eta&+\int_{-r}^{d0}G(u)\{x(t+u)-x(u)\}\,du\\&+\int_0^t\int_{-r}^{0}F(u)x(s+u)\,du\,ds+\int_0^t f(u)\,du.\end{aligned} \tag{3.4}$$

If $(\eta,\varphi)\in\mathbb{R}^n\times L_p$ and $f\in L_p^{\mathrm{loc}}(0,+\infty)$, then it follows from Lemma 2.6 that there exists a unique solution $x=x(\cdot;\eta,\varphi,f)$ to (3.1)–(3.2). In particular, there is a unique pair $(y(\cdot;\eta,\varphi,f),x(\cdot;\eta,\varphi,f))$ satisfying (3.3)–(3.4) with $x(s;\eta,\varphi,f)=\varphi(s)$ a.e. on $[-r,0]$. If $t_1>0$ and $0\leq t\leq t_1$, then Lemma 2.7 implies that the mapping $(\eta,\varphi,f)\to x_t(\cdot;\eta,\varphi,f)$ from $\Gamma=\mathbb{R}^n\times L_p\times L_p(0,t_1)$ into L_p is continuous. Since $x_t\in L_p$ it follows that $y(t;\eta,\varphi,f)$ is continuous in t and for each fixed $t\in[0,t_1]$ the mapping $(\eta,\varphi,f)\to y(t;\eta,\varphi,f)$ from Γ into $\mathbb{R}^n$ is also continuous. Consequently, we have the following result.

LEMMA 3.1. *Let $t_1>0$ be fixed and assume that $(\eta^N,\varphi^N,f^N)\to(\eta,\varphi,f)$ in Γ. Then for each $t\in[0,t_1]$*

$$\lim_{N\to\infty}\|x_t(\cdot;\eta^N,\varphi^N,f^N)-x_t(\cdot;\eta,\varphi,f)\|_{L_p}=0,$$

and

$$\lim_{N\to\infty}\|y(t;\eta^N,\varphi^N,f^N)-y(t;\eta,\varphi,f)\|_{\mathbb{R}^n}=0.$$

Let $\mathcal{A}$ be defined by (2.1)–(2.2) and consider the abstract Cauchy problem in $\mathbb{R}^n\times L_p$

$$\dot{z}(t)=\mathcal{A}z(t)+(f(t),0),\qquad t>0, \tag{3.5}$$

with initial data

$$z(0)=(\eta,\varphi). \tag{3.6}$$

Since $L\in\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and $D\in\mathcal{B}(C,\mathbb{R}^n)$ is atomic at $s=0$, it follows from Theorem 2.3 that $\mathcal{A}$ generates a C_0-semigroup $\{\mathcal{S}(t)\}_{t\geq 0}$ on $\mathbb{R}^n\times L_p$. Consequently, we may define the *mild solution* of (3.5)–(3.6) $z=z(\cdot;\eta,\varphi,f)$, by

$$z(t;\eta,\varphi,f)=\mathcal{S}(t)(\eta,\varphi)+\int_0^t\mathcal{S}(t-s)(f(s),0)\,ds. \tag{3.7}$$

Note that in general z defined by (3.7) may not be a solution to (3.5)–(3.6) (in the classical sense). However, we do have a basic equivalence between z and solutions of (3.1)–(3.2) (see [2], [21] for similar results).

THEOREM 3.1. *Let $t_1>0$ be fixed. If $(\eta,\varphi)\in\mathbb{R}^n\times L_p$ and $f\in L_p(0,t_1)$, then*

$$z(t;\eta,\varphi,f)=(y(t;\eta,\varphi,f),x_t(\cdot;\eta,\varphi,f)), \tag{3.8}$$

for all $0\leq t\leq t_1$, where z is defined by (3.7) *and (y,x) is the solution to the system* (3.3)–(3.4).

Proof. Pick $(\eta^N,\varphi^N)\in\mathcal{D}(\mathcal{A})$ and f^N to be continuously differentiable for $N=1,2,\cdots$, such that $(\eta^N,\varphi^N,f^N)\to(\eta,\varphi,f)$ in Γ. For each N, $z(\cdot;\eta^N,\varphi^N,f^N)$ defined by (3.7) is a continuously differentiable solution of (3.5)–(3.6) (see [27]). However, in this case Lemma 2.8 implies that $z(t;\eta^N,\varphi^N,f^N)=(Dx_t(\cdot;\eta^N,\varphi^N,f^N),\ x_t(\cdot;\eta^N,\varphi^N,f^N))=(y(t;\eta^N,\varphi^N,f^N),\ x_t(\cdot;\eta^N,\varphi^N,f^N))$ for all $N=1,2,\cdots$. Consequently, the continuity of z in (η,φ,f) and Lemma 3.1 yield

$$\begin{aligned} z(t;\eta,\varphi,f)&=\lim_{N\to\infty}z(t;\eta^N,\varphi^N,f^N)\\ &=\lim_{N\to\infty}\left(y(t;\eta^N,\varphi^N,f^N),\,x_t(\cdot;\eta^N,\varphi^N,f^N)\right)\\ &=(y(t,\eta,\varphi,f),\,x_t(\cdot,\eta,\varphi,f)), \end{aligned}$$

and the theorem is established. □

For each $t\in[0,t_1]$, define the operator $\mathcal{G}(t)\colon L_p(0,t_1)\to\mathbb{R}^n\times L_p$ by

$$\mathcal{G}(t)f=\int_0^t\mathcal{S}(t-s)(f(s),0)\,ds, \tag{3.9}$$

and define $\mathcal{F}\colon L_p(0,t_1)\to C([0,t_1];\mathbb{R}^N\times L_p)$ by

$$[\mathcal{F}f](t)=\mathcal{G}(t)f. \tag{3.10}$$

The following result may be established by trivial modifications of the proof for the retarded case (see [21]).

THEOREM 3.2. *The operator $\mathcal{F}$ is compact. In particular, for each $t\in[0,t_1]$ the operator $\mathcal{G}(t)$ is compact.*

Remark. 3.1. If $f^N\to f$ weakly in $L_p(0,t_1)$, then it follows that $\mathcal{G}(t)f^N\to\mathcal{G}(t)f$ uniformly for $t\in[0,t_1]$. Since the mild solutions (3.7) may be written as

$$z(t;\eta,\varphi,f)=\mathcal{S}(t)(\eta,\varphi)+\mathcal{G}(t)f, \tag{3.11}$$

it follows that $z(t;\eta^N,\varphi^N,f^N)$ converges *strongly* to $z(t;\eta,\varphi,f)$ if $(\eta^N,\varphi^N)\to(\eta,\varphi)$ and $f^N\to f$ weakly. Moreover, the convergence is uniform for $t\in[0,t_1]$. Because of this feature and the equivalence (3.8), the formulation (3.5)–(3.7) provides a particularly nice framework for the study of approximation techniques for optimal control and identification of NFDEs (see [2], [3], [11], [13] for similar results for retarded systems and [21] for neutral systems).

4. A nonatomic D operator. In this section we show that the sufficient condition that D be atomic (see Theorem 2.3) can in some cases be relaxed. In particular we define $D\varphi \equiv \int_{-1}^{0} \varphi(s)|s|^{-\alpha}\,ds$ with $0<\alpha<1$. It is to be noted that D is bounded on C for all $\alpha\in(0,1)$ and is bounded on L_p if $p>1/(1-\alpha)$; however, D is unbounded yet densely defined on L_p when $p\le 1/(1-\alpha)$. For simplicity we take $L\equiv 0$ and consider the scalar NFDE

$$\frac{d}{dt}Dx_t=0, \quad t>0, \qquad x_0=\varphi. \tag{4.1}$$

THEOREM 4.1. (i) *For each $\varphi\in C$, the initial value problem*

$$Dx_t=D\varphi, \quad t\ge 0, \qquad x_0=\varphi \tag{4.2}$$

has a unique continuous solution $x(\cdot;\varphi)$ on $[0,\infty)$. The family of operators $T(t)\varphi = x_t(\cdot;\varphi)$; $t\ge 0$ defines a C_0-semigroup on C.

(ii) *If $p<1/(1-\alpha)$ (i.e., $\alpha>1-1/p$) then for each $(\eta,\varphi)\in\mathbb{R}\times L_p$ the initial value problem*

$$Dx_t=\eta, \quad t\ge 0, \qquad x_0=\varphi, \tag{4.3}$$

has a unique solution $x(\cdot;\varphi)$ defined a.e. on $[0,+\infty)$. Moreover, the family of operators $S(t)(\eta,\varphi)=(Dx_t,x_t)=(\eta,x_t(\cdot;\varphi))$, $t\ge 0$ defines a C_0-semigroup on $\mathbb{R}\times L_p$.

(iii) *If $p\ge 1/(1-\alpha)$ (i.e., $\alpha\le 1-1/p$) then* (4.3) *has a unique solution for (η,φ) in a dense subset of $\mathbb{R}\times L_p$. However, the family of operators $S(t)$ defined above fails to define a C_0-semigroup on $\mathbb{R}\times L_p$.*

Note that since $L\equiv 0$, the semigroup of part (ii) of Theorem 4.1 coincides with the semigroup of Theorem 2.2 (see (2.9)). The above theorem illustrates that the choice of the phase space for which (4.1) is well posed is dependent (in a sensitive way) on the singularity of the operator D. The case $\alpha=\frac{1}{2}$, $p=2$ is of particular interest in that (4.1) arises in some models of aerodynamics [3]. In this case we have D densely defined and unbounded but $S(t)$ fails to define a C_0-semigroup on $\mathbb{R}\times L_2$.

The following lemmas will be needed for the proof of Theorem 4.1. The first, stated without proof, is a well-known result concerning the solution of Abel's integral equation (see Hochstadt [20]).

LEMMA 4.1. *Let $f\in L_\infty(0,1)$. Then $x(\cdot)$ satisfies*

$$\int_0^t x(s)|t-s|^{-\alpha}\,ds=f(t) \quad \text{a.e. on } (0,1) \tag{4.4}$$

if and only if

$$\int_0^t x(s)\,ds=\frac{\sin a\pi}{\pi}\int_0^t f(s)|t-s|^{\alpha-1}\,ds \quad \text{on } [0,1]. \tag{4.5}$$

LEMMA 4.2. *Let $\varphi\in C$, $\eta\in\mathbb{R}$. Then* (4.3) *has the unique integrable solution*

$$\begin{aligned} x(t)=&\frac{\sin(\alpha\pi)}{\pi}\int_{-1}^{0}\frac{1}{t-s}\left|\frac{t}{s}\right|^{\alpha}\varphi(s)\,ds \\ &+\frac{\sin(\alpha\pi)}{\pi}\int_0^t\frac{(t-s)^{\alpha-1}}{(t-s)+1}\varphi(s-1)\,ds+[\eta-D\varphi]t^{\alpha-1}, \qquad 0<t\le 1. \end{aligned} \tag{4.6}$$

Proof. Given $(\eta,\varphi)\in\mathbb{R}\times C$ we define the continuous function f on $[0,1]$ by

$$f(t)=\int_{-1}^{0}\left[|s|^{-\alpha}-|t-s|^{-\alpha}\right]\varphi(s)\,ds+\int_0^t(t-u+1)^{-\alpha}\varphi(u-1)\,du+[\eta-D\varphi]$$

and note that (4.3) becomes (4.4) with f defined as above. Consequently, our proof is complete if we can show that the right-hand side of (4.5) is differentiable where f is as defined above. Changing the order of integration, and making several changes of variables we have

$$\begin{aligned}
&\frac{d}{dt}\int_0^t\int_{-1}^0\left(|u|^{-\alpha}-|s-u|^{-\alpha}\right)\varphi(u)\,du\,|t-s|^{\alpha-1}\,ds\\
&\qquad=\frac{d}{dt}\int_{-1}^0\int_0^t\left(|u|^{-\alpha}-|s-u|^{-\alpha}\right)(t-s)^{\alpha-1}\,ds\varphi(u)\,du\\
&\qquad=\frac{d}{dt}\int_{-1}^0\int_0^t\left(\frac{1}{|u|^{\alpha}}-\frac{1}{|t-w-u|^{\alpha}}\right)w^{\alpha-1}\,dw\,\varphi(u)\,du\\
&\qquad=\int_{-1}^0\int_0^t\alpha(t-w-u)^{-\alpha-1}w^{\alpha-1}\,dw\,\varphi(u)\,du\\
&\qquad=\int_{-1}^0\frac{\alpha}{t-u}\int_0^{t/(t-u)}(1-\theta)^{-\alpha-1}\theta^{\alpha-1}\,d\theta\,\varphi(u)\,du\\
&\qquad=\int_{-1}^0\frac{\alpha}{t-u}\int_0^{-t/u}\gamma^{\alpha-1}\,d\gamma\,\varphi(u)\,du\\
&\qquad=\int_{-1}^0\frac{1}{(t-u)}\left|\frac{t}{u}\right|^{\alpha}\varphi(u)\,du.
\end{aligned}$$

For the second term of f, we have that

$$\begin{aligned}
&\frac{d}{dt}\int_0^t\int_0^s(s-u+1)^{-\alpha}\varphi(u-1)\,du|t-s|^{\alpha-1}\,ds\\
&\qquad=\frac{d}{dt}\int_0^t\int_u^t(s-u+1)^{-\alpha}(t-s)^{\alpha-1}\,ds\,\varphi(u-1)\,du\\
&\qquad=\frac{d}{dt}\int_0^t\int_0^{t-u}(\theta+1)^{-\alpha}(t-\theta-u)^{\alpha-1}\,d\theta\,\varphi(u-1)\,du\\
&\qquad=\frac{d}{dt}\int_0^t b(t-u)\varphi(u-1)\,du,
\end{aligned}$$

where

$$b(v)\equiv\int_0^v\frac{1}{(v-\theta)^{1-\alpha}}\,\frac{1}{(\theta+1)^{\alpha}}\,d\theta.$$

The commutative property for convolutions together with a change of variables yield

$$b(v)=\int_0^{1/(1+v)}(w)^{\alpha-1}(1-w)^{-\alpha}\,dw.$$

Using this representation for b we obtain

$$\frac{d}{dt}\int_0^t b(t-u)\varphi(u-1)\,du=\int_0^t\frac{(t-u)^{\alpha-1}}{1+t-u}\varphi(u-1)\,du.$$

Finally, we note that

$$\frac{d}{dt}\int_0^t(t-u)^{\alpha-1}[\eta-D\varphi]\,du=[\eta-D\varphi]t^{\alpha-1}.$$

□

Remark 4.1. From (4.6) and $0<\alpha<1$ it is clear that (4.3) does not have a continuous solution on $[0,1]$ unless $\eta-D\varphi=0$. The condition $p<1/(1-\alpha)$ is necessary for (4.3) to be well posed on $\mathbb{R}\times L_p$.

Proof of Theorem 4.1(i). An application of Lemma 4.2 with $\eta=D\varphi$ yields the unique integrable solution $x(t)$, $0<t\leq 1$, given by (4.6) with $\eta-D\varphi=0$. The first and second terms of the representation (4.6) for $x(t)$ are clearly continuous on $(0,1]$ and $[0,1]$, respectively. In order to establish the continuity of $x(t)$ at $t=0$ we note that

$$\begin{aligned}\lim_{t\to 0^+}\int_{-1}^{0}\frac{1}{t-s}\left|\frac{t}{s}\right|^{\alpha}\varphi(s)\,ds &= \lim_{t\to 0^+}\int_0^{1/t}(1+\theta)^{-1}\theta^{-\alpha}\varphi(-t\theta)\,d\theta\\ &=\int_0^{\infty}(1+\theta)^{-1}\theta^{-\alpha}\varphi(0)\,d\theta\\ &=\varphi(0)\frac{\pi}{\sin(\alpha\pi)}.\end{aligned}$$

Therefore a unique continuous solution x of (4.2) exists on $[0,1]$. The "method of steps" is employed to obtain a unique continuous solution on $[0,+\infty)$. The boundedness of the usual solution operators $T(t)$; $t\geq 0$ follows from (4.6) while the semigroup properties of $T(t)$ and the strong continuity in t follow by well-known arguments.

Remark 4.2. A modification of Levinson [25, Lemma 2.1] yields the existence of a constant m (dependent only on $\|\varphi\|$ so that $|x(t;\varphi)-x(s;\varphi)|\leq m|t-s|^{\alpha}$ for $0\leq s,t\leq 1$). As an immediate consequence of this inequality we have that the semigroup operators $T(t)$ are compact for $t\geq 1$.

Proof of Theorem 4.1(ii). By Theorem 2.2 it suffices to show that $\mathcal{A}(\eta,\varphi)=(0,\dot{\varphi})$ with $\mathcal{D}(\mathcal{A})=\{(\eta,\varphi)\in\mathbb{R}\times L_p/\varphi\in W^{1,p},\ D\varphi=\eta\}$ defines a C_0-semigroup on $\mathbb{R}\times L_p$. We proceed as in §2 and verify the conditions of Phillips' theorem [9], [27]. Recall that $\alpha>1-1/p$, thus D is unbounded and the density of $\mathcal{D}(\mathcal{A})$ follows immediately. Lemma 2.3 yields $\{\lambda\in\mathbb{R}|\lambda>0\}\subset\rho(\mathcal{A})$.

To complete the proof we need only show that the Cauchy problem (4.2) with $\varphi\in W^{1,p}$ has a unique solution $x_t(\cdot,\varphi)=x_t(\varphi)$ which is continuously differentiable as a function of t into L_p. Since $W^{1,p}\subset C$ part (i) of Theorem 4.1 proves the existence of the continuous solution (4.6) (with $D\varphi=\eta$) on $[0,\infty)$. We proceed to show $\frac{d}{dt}x_t(\varphi)=\dot{x}_t(\varphi)\in L_p$ for $0\leq t\leq 1$ and the desired result will then follow by the method of steps. Applying Lemma 2.4 to the second term in (4.6) we note that this convolution lies in $W^{1,p}[0,1]$ when $\varphi\in W^{1,p}$. The following lemma completes our proof of part (ii) of Theorem 4.1.

LEMMA 4.3. *Let* $\varphi\in W^{1,p}$ *and define* $h(t)=\int_{-1}^{0}1/(t-s)|t/s|^{\alpha}\varphi(s)\,ds$ *for* $0\leq t\leq 1$, *then* $\dot{h}(t)\in L_p(0,1)$ *provided* $\alpha>1-1/p$.

Proof. The change of variables $\theta=-s/t$ gives $h(t)=\int_0^{1/t}(1+\theta)^{-1}\theta^{-\alpha}\varphi(-\theta t)\,d\theta$; therefore we have

$$\dot{h}(t)=-\frac{t^{\alpha-1}}{t+1}\varphi(-1)-\int_0^{1/t}(1+\theta)^{-1}\theta^{-\alpha+1}\dot{\varphi}(-\theta t)\,d\theta$$

or, equivalently,

$$\begin{aligned}\dot{h}(t)=-\frac{t^{\alpha-1}}{t+1}\varphi(-1)&-\int_0^{t}\frac{1}{t+s}\left(\frac{s}{t}\right)^{1-\alpha}\dot{\varphi}(-s)\,ds\\ &-\int_t^{1}\frac{1}{t+s}\left(\frac{s}{t}\right)^{1-\alpha}\dot{\varphi}(-s)\,ds,\qquad t\in[0,1].\end{aligned}\tag{4.7}$$

For convenience we denote the three terms on the right side of (4.7) by $I_1(t)$, $I_2(t)$ and $I_3(t)$, respectively. Since $\alpha>1-1/p$, $I_1\in L_p(0,1)$. Since $1/p+\alpha-1<1$ an application of Lemma 7.23 of Adams [1] to I_2 yields

$$\int_0^1|I_2(t)|^p dt\le\int_0^1\left(t^{\alpha-2}\int_0^t s^{1-\alpha}|\dot\varphi(-s)|\,ds\right)^p dt\le\frac{1}{\left(2-\frac{1}{p}-\alpha\right)^p}\int_0^1|\dot\varphi(-s)|^p ds.$$

As for I_3, we take arbitrary $\psi\in L_q$ and consider

$$\begin{aligned}\left|\int_0^1\psi(t)I_3(t)\,dt\right|&=\left|\int_0^1\int_0^s\psi(t)\frac{s^{1-\alpha}}{t+s}t^{\alpha-1}\,dt\,\dot\varphi(-s)\,ds\right|\\&\le\int_0^1 s^{-\alpha}|\dot\varphi(-s)|\int_0^s|\psi(t)|t^{\alpha-1}\,dt\,ds\\&\le\left\{\int_0^1\left(s^{-\alpha}\int_0^s|\psi(t)|t^{\alpha-1}\,dt\right)^q ds\right\}^{1/q}\left\{\int_0^1|\dot\varphi(-s)|^p ds\right\}^{1/p}\\&\le\frac{1}{\left(\frac{1}{p}+\alpha-1\right)}\left\{\int_0^1|\psi(t)|^q dt\right\}^{1/q}\left\{\int_0^1|\dot\varphi(-s)|^p ds\right\}^{1/p},\end{aligned}$$

where the last inequality is a consequence of Adams [1, Lemma 7.23]. Thus, $I_3\in L_p(0,1)$ with norm not larger than $[1/((1/p)+\alpha-1)](\int_0^1|\dot\varphi(-s)|^p\,ds)^{1/p}$.

Proof of Theorem 4.1(iii). This follows immediately from Lemma 4.2 and Remark 4.1.

Remark 4.3. If $\alpha=0$ in the above example then local L_1 solutions of (4.1) coincide with those of

$$x(t)=x(t-1),\quad t>0,\qquad x_0=\varphi. \tag{4.8}$$

If $\varphi\in C$, (4.8) has a continuous solution on $(0,\infty)$ if and only if $\varphi(0)=\varphi(-1)$. Thus the problem (4.8) for $\varphi\in C$ is in general, ill-posed. Similarly if $\varphi\in L_p$ then

$$\int_{-1}^0 x(t+s)\,ds=\eta,\quad t>0,\qquad x_0=\varphi,$$

has a unique solution on $(0,+\infty)$; however, $\mathcal{A}(\eta,\varphi)\equiv(0,\dot\varphi)$ with domain $\{(\eta,\varphi)\in\mathbb{R}\times L_p|D\varphi=\eta\}$ fails to generate a C_0-semigroup on $\mathbb{R}\times L_p$ since D is a bounded functional on L_p. Finally, if $x(t;\varphi)$ denotes the solution of (4.8) for $t\in\mathbb{R}$ and $\varphi\in L_p$, then the solution operator $U(t)\varphi=x_t(\cdot;\varphi)$ defines a strongly continuous group of isometries on L_p, $1\le p<\infty$.

5. Concluding remarks. The results in §2 extend and refine the paper by Delfour [10]. We established that if $\mathcal{A}$ defined by (2.1)–(2.2) generates a C_0-semigroup on $\mathbb{R}^n\times L_p$, then both L and D belong to $\mathcal{B}(W^{1,p},\mathbb{R}^n)$ and D cannot belong to $\mathcal{B}(L_p,\mathbb{R}^n)$. In order to obtain general sufficient conditions, we imposed stronger conditions on D; namely that $D\in\mathcal{B}(C,\mathbb{R}^n)$ and has an atom at $s=0$. However, in §4 we gave an example of a singular integral equation where $D\in\mathcal{B}(C;\mathbb{R}^n)$ was nonatomic and yet $\mathcal{A}$ generated a C_0-semigroup on the product space $\mathbb{R}^n\times L_p$. Consequently, the assumption that D have an atom is not necessary. These observations lead to the following open question: Is there a set of conditions on L and D that are both necessary and sufficient for $\mathcal{A}$ to generate a C_0-semigroup on $\mathbb{R}^n\times L_p$? It is clear that these conditions must be

stronger than the necessary conditions given by Theorems 2.1 and 2.2 yet weaker than the sufficient condition required in Theorem 2.3.

The equivalence between the NFDEs and the abstract integral equation (3.7) provides an excellent framework to study approximation techniques for neutral systems. We shall investigate particular approximation schemes in a future paper. However, we note that Kappel [21] has already made use of this basic idea in his study of approximation schemes for certain NFDEs.

Perhaps one of the most interesting aspects of this paper concerns the integral equation discussed in §4. For the case where the integral equation generates a dynamical system on $\mathbb{R}^n \times L_p$, there is again the possibility of using general approximation schemes for well-posed problems to study numerical methods for these singular type integral equations. This problem certainly seems worthy of further study.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), 169–208.

[3] H. R. BANKS, J. A. BURNS AND E. M. CLIFF, *Parameter estimation and identification for systems with delays*, SIAM J. Control Optim., 19 (1981), pp. 791–828.

[4] JU. BORISOVIČ AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded arguments*, Soviet Math. Dokl., 10 (1969), pp. 401–405.

[5] R. C. BROWN AND A. M. KRALL, *Ordinary differential operators under Stieltjes boundary conditions*, Trans. Amer. Math. Soc., 198 (1974), pp. 73–92.

[6] J. A. BURNS AND T. L. HERDMAN, *Adjoint semigroup theory for a class of functional differential equations*, this Journal, 5 (1976), pp. 729–745.

[7] J. A. BURNS, T. L. HERDMAN AND H. W. STECH, *Differential-boundary operators and associated neutral functional differential equations*, Rocky Mountain J. Math., to appear.

[8] ———, *The Cauchy problem for linear functional differential equations*, Integral and Functional Differential Equations, Lecture Notes in Pure and Applied Mathematics Vol. 67, T. L. Herdman, S. M. Rankin, and H. W. Stech, eds. Marcel Dekker, New York, 1981, pp. 139–149.

[9] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[10] M. C. DELFOUR, *The largest class of hereditary systems defining a C_0-semigroup on the product space*, Canad. J. Math, 32 (1980), pp. 969–978.

[11] ———, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Applied Math. Optim, 3 (1977), pp. 101–162.

[12] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298–328.

[13] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[14] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[15] ———, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[16] D. HENRY, *Linear autonomous functional differential equations of neutral type in Sobolev space $W_2^{(1)}$*, unpublished manuscript.

[17] R. HERSH AND T. KATO, *High-accuracy stable difference schemes for well-posed initial value problems*, SIAM J. Numer. Anal., 16 (1979), pp. 670–694.

[18] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis*, Vol. III, Springer-Verlag, New York, 1965.

[19] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, New York, 1969.

[20] H. HOSCHSTADT, *Integral Equations*, Pure and Applied Mathematics, Wiley-Interscience, New York, 1973.

[21] F. KAPPEL, *Approximation of neutral functional differential equations in the state space $\mathbb{R}^n \times L^p$*, preprint.

[22] F. KAPPEL AND K. KUNISCH, *Spline approximations for neutral functional differential equations*, SIAM J. Numer. Anal., 18 (1981), pp. 1058–1080.

[23] E. KREYSZIG, *Introduction to Functional Analysis with Applications*, John Wiley, New York, 1978.

[24] K. KUNISCH, *Approximation schemes for the linear-quadratic optimal control problem associated with delay-equations*, SIAM J. Control Optim., 20 (1982), pp. 506–540.

[25] N. LEVINSON, *A nonlinear Volterra equation arising in the theory of superfluidity*, J. Math. Analysis and Applic., 1 (1960), pp. 1–11.

[26] A. MANITIUS, *Optimal control of hereditary systems*, in Control Theory and Topics in Functional Analysis, Vol. III, IAEA, Vienna, 1976.

[27] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Math. Dept. Lecture Notes, Vol. 10, Univ. Maryland, College Park, 1974.

[28] D. C. REBER, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 193–232.

[29] H. F. TROTTER, *Approximations of semigroups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.

[30] R. B. VINTER, *On a problem of Zabczyk concerning semigroups generated by operators with non-local boundary conditions*, Publication 77/8, (1977), Department of Computing and Control, Imperial College of Science and Technology, London.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0008 $01.25/0

LOCAL EXISTENCE THEOREMS FOR NONLINEAR DIFFERENTIAL EQUATIONS*

NORIMICHI HIRANO†

Abstract. Let X be a real Banach space and $U \subset X$ be an open set. In this paper, we consider the initial value problem:

$$\frac{du}{dt} + Au \ni G(u) \quad (0 \le t \le T), \qquad u(0) = u_0,$$

where G: $C(0, T; U) \to L^\infty(0, T; X)$ is a given mapping and A is a given m-accretive set such that $(I + \lambda A)^{-1}$ is compact for all $\lambda > 0$. We also study the problem above under the assumption that A is weakly closed and G is a pointwise defined function.

1. Introduction. Let X be a real Banach space with norm $\| \ \|$, $U \subset X$ be an open set. In this paper we study the existence of local solutions to the initial value problem:

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) \ni g(t, u(t)), \qquad 0 \le t \le T, \\ &u(0) = u_0. \end{aligned} \tag{1.1}$$

where $A \subset X \times X$ is m-accretive, g is a continuous mapping from $[0, a] \times U$ into X and $u_0 \in \overline{D(A)} \cap U$.

Also, we study the generalized problem of (1.1):

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) \ni G(u)(t), \qquad 0 \le t \le T, \\ &u(0) = u_0. \end{aligned} \tag{1.2}$$

where $A \subset X \times X$ is m-accretive, G is a mapping from $C(0, T; X)$ into $L^\infty(0, T; X)$ and $u_0 \in \overline{D(A)} \cap U$.

It is well known that if A is linear and g satisfies a local Lipschitz condition, then local solutions of (1, 1) exist (see [9]). Also it is known that if A is linear, $g(t, \cdot)$ is dissipative for all $0 \le t \le a$ and g maps bounded sets into bounded sets, then a unique solution of (1.1) exists (see [7]). In the case in which G satisfies an appropriate Lipschitz condition, the initial value problem (1.2) was studied by Crandall and Nohel [8], while Pazy [12] studied the problem (1.1) in the case in which A is linear and generates a semigroup of compact operators. Recently Vrabie [13] extended Pazy's result and established a local existence theorem for the problem (1.2) in the case in which G is continuous and A generates a semigroup of compact operators.

In §2, we consider the problem (1.2) in the case in which $(I + \lambda A)^{-1}$ is compact for all $\lambda > 0$. The proof of the result in §2 is given in §3. In §4 we study the problem (1.1) under the assumption of weak closedness of A. Throughout the rest of the present paper, X is a real Banach space with norm $\| \cdot \|$ and X^* is its dual space with the corresponding norm $\| \cdot \|_*$. "$\to$" and "$\rightharpoonup$" indicate strong and weak convergence, respectively. $\overline{D}$ denotes the closure of D and $\overline{\text{co}}D$ denotes the closed convex hull of D. The normalized duality mapping J of X into X^* is given by

$$J(x) = \{x^* \in X^* : (x, x^*) = \|x\| \|x^*\| = \|x\|^2\} \tag{1.3}$$

*Received by the editors June 22, 1981, and in revised form October 30, 1981.

†Department of Information Sciences, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo, Japan.

for $x \in X$. For each $\langle x,y\rangle \in X\times X$, define

$$\langle x,y\rangle_+ = \sup\{(x,j): j\in J(x)\}. \tag{1.4}$$

If X^* is assumed to be a uniformly convex Banach space, then J is single valued and uniformly continuous on bounded sets. If $A \subset X\times X$ and $x\in X$, we define $Ax=\{y\in X: \langle x,y\rangle \in A\}$, $D(A)=\{x\in X: Ax\neq \emptyset\}$, $R(A)=\cup\{Ax: x\in D(A)\}$, and $|Ax|=\inf\{\|y\|: y\in Ax\}$. A subset $A\subset X\times X$ is said to be weakly closed if $\langle x_\lambda, y_\lambda\rangle \in A$, $x_\lambda \to x$, and $y_\lambda \to y$ implies $\langle x,y\rangle \in A$. For $x\in X$ and $r>0$, we denote by $B(x,r)$ the closed ball with center x and radius r. Let $T>0$. $C(0,T;X)$ denotes the space of all continuous functions from $[0,T]$ into X. X_w denotes the space X equipped with the weak topology, and $C(0,T;X_w)$ denotes the topological vector space of functions defined on $[0,T]$ with values in X and continuous in the weak topology. $L^p(0,T;X)$ denotes the space of all X-valued p-integrable functions defined almost everywhere on $[0,T]$. For each $u\in L^1(0,T;X)$, we denote by $\mathrm{var}(u;[0,t])$ the variation of u on $[0,t]$.

We restate here some results of the theory of evolution equations which are used in the present paper. Let $A\subset X\times X$ be an m-accretive set and $f\in L^1(0,T;X)$. A function $u: [0,T]\to X$ is called a strong solution of the initial value problem:

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) \ni f(t), \qquad 0\le t\le T,\\ &u(0)=u_0, \end{aligned} \tag{1.5}$$

if u is differentiable almost everywhere on $[0,T]$, absolutely continuous, and satisfies $u(0)=u_0$ and $u'(t)+Au(t)\ni f(t)$ almost everywhere on $[0,T]$. An integral solution of (1.5) is a function $v: [0,T]\to X$ such that v is continuous on $[0,T]$, $v(0)=u_0$ and satisfies the inequality:

$$\|v(t)-x\|^2 \le \|v(s)-x\|^2 + 2\int_s^t \langle f(s)-y, v(s)-x\rangle_+ \, ds \tag{1.6}$$

for all $\langle x,y\rangle \in A$ and $0\le s\le t\le T$. Every strong solution of (1.5) is also an integral solution of (1.5). If u and v are integral solutions of (1.5) corresponding to $h\in L^1(0,T;X)$ and $g\in L^1(0,T;X)$, respectively, then

$$\|u(t)-v(t)\|^2 \le \|u(s)-v(s)\|^2 + 2\int_0^t \langle h(\tau)+g(\tau), u(\tau)-v(\tau)\rangle_+ \, d\tau. \tag{1.7}$$

It is known that the initial value problem (1.5) has a unique integral solution [4], [5]. In particular, if A is continuous with $D(A)=X$ and f is continuous, then (1.5) has a unique strong solution.

2. The main result. In [13], Vrabie established a local existence theorem for integral solutions to the initial value problem:

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) \ni G(u)(t), \qquad 0\le t\le T,\\ &u(0)=u_0, \end{aligned} \tag{2.1}$$

under the condition that A generates a nonlinear semigroup $S(t)$: $\overline{D(A)}\to\overline{D(A)}$ with $S(t)$ compact for all $t>0$, while $U\subset X$ is open and G: $C(0,a;U)\to L^\infty(0,a;X)$ is continuous. Concerning Vrabie's result, we restate the necessary and sufficient condition for $S(t)$ to be compact for all $t>0$.

THEOREM (Brézis [6]). *Let $S(t)$ be the semigroup generated by A. Then the following properties are equivalent.*

(i) *For each $t>0$, $S(t)$ is compact, i.e., $S(t)$ maps bounded sets of $\overline{D(A)}$ into relatively compact sets of X.*

(ii) (iia) *For each $\lambda>0$, $(I+\lambda A)^{-1}$ is compact, i.e., $(I+\lambda A)^{-1}$ maps bounded sets of X into compact sets of X;*

(iib) *for each bounded set B in $\overline{D(A)}$, the family $\{S(\cdot)x\colon R^+\to X\colon x\in B\}$ is equicontinuous.*

The condition (iib) restricts the applicability of Vrabie's result essentially to the "parabolic" problem. In this section, we consider the problem (1.2) under the condition (iia) which applies to a broad class of equations.

THEOREM 2.1. *Let X be a real Banach space and $A\subset X\times X$ be an m-accretive set such that $(I+\lambda A)^{-1}$ is compact for all $\lambda>0$. Let $U\subset X$ be an open set and $G\colon C(0,a;U)\to C(0,a;X)$ satisfy the following conditions:*

(C1) *G; $C(0,a;U)\to C(0,a;X)$ is continuous.*

(C2) *There exists a function $k\colon(0,\infty)\to(0,\infty)$ such that,*

$$\operatorname{var}(G(u)\colon[0,t])\le k(d)(1+\operatorname{var}(u\colon[0,t])), \tag{2.2}$$

whenever $u\in C(0,a;U)$ is of bounded variation and $\|u(t)\|\le d$ for all $0\le t\le a$.

Then for each $u_0\in D(A)\cap U$, there exists $T\in(0,a]$ such that (1.2) has an integral solution on $[0,T]$.

COROLLARY 2.1. *Let X be a reflexive Banach space and A, U be as in Theorem 2.1. Let $G\colon C(0,a;U)\to L^\infty(0,a;X)$ satisfy (Ç2) and the following condition:*

(C1′) *$G\colon C(0,a;U)\to L^\infty(0,a;X)$ is continuous.*

Then for each $u_0\in D(A)\cap U$, there exists $T\in(0,a]$ such that (1.2) has a strong solution on $[0,T]$.

COROLLARY 2.2. *Let X, A, U and G be as in Theorem 2.1. In addition, assume that A is linear. Then for each $u_0\in D(A)\cap U$ there exists $T\in(0,a]$ such that (1.2) has a strong solution $u\in C(0,T;U)$ and it satisfies*

$$u(t)=S(t)u_0+\int_0^t S(t-s)G(u)(s)\,ds,$$

for all $0\le t\le T$, where $\{S(t)\}$ is the semigroup generated by A.

Remark. Condition (C2) was introduced by Crandall and Nohel [8] to guarantee the Lipschitz continuity of solutions of (1.2). In Theorem 2.1, we need (C2) essentially to show the existence of the solutions of (1.2).

3. Proof of Theorem 2.1. First we define operators

$$J_n=(I+n^{-1}A)^{-1},\qquad A_n=n(I-J_n),\qquad n=1,2,\cdots.$$

It is well known that the J_n and A_n are single valued and uniformly Lipschitz continuous with

$$\|J_nx-J_ny\|\le\|x-y\|,\qquad \|A_nx-A_ny\|\le 2n\|x-y\|,$$

for all $x,y\in X$. Also it is known that for each $n\ge 1$, A_n is m-accretive and $\|A_nx\|\le|Ax|$ for all $x\in D(A)$. Let $\{S_n(t)\}_{t>0}$ be the semigroup generated by A_n. Then for each $x\in\overline{D(A)}$,

$$S_n(t)x\to S(t)x\quad\text{as } n\to\infty$$

uniformly on a bounded interval (see [3] for details). Let $u_0 \in D(A) \cap U$. Next we choose positive numbers M, r and T such that $B(u_0, r) \subset U$, $k(r + \|u_0\|)T < 1$ and

$$\|G(u)(t)\| \leq M \quad \text{a.e. on } [0, T], \tag{3.1}$$

for all $u \in C(0, T; U)$ with $u(t) \in B(u_0, r)$ for all $0 \leq t \leq T$, and in addition,

$$TM + \|S(t)u_0 - u_0\| \leq r \tag{3.2}$$

for all $0 \leq t \leq T$. Since G is continuous and $S(t)$ is continuous at the origin, it is possible to choose such positive numbers M, r and T. Now, we set

$$K = \{u \in C(0, T; U) : u(t) \in B(u_0, r) \text{ for all } 0 \leq t \leq T\}, \tag{3.3}$$

$$K(V) = \{u \in K : \operatorname{var}(u : [0, T]) \leq V\} \tag{3.4}$$

for each $V > 0$, and define an operator Q and a sequence of operators $\{Q_n\}$ by the method employed in [13]. Let $n \geq 1$ and $u \in K$. Then the initial value problem

$$\begin{aligned} &\frac{dv(t)}{dt} + A_n v = G(u)(t), \qquad 0 \leq t \leq T, \\ &v(0) = u_0 \end{aligned} \tag{3.5}$$

has a unique strong solution $v \in C(0, T; X)$. Also the initial value problem

$$\begin{aligned} &\frac{dw(t)}{dt} + Aw(t) \ni G(u)(t), \qquad 0 \leq t \leq T, \\ &w(0) = u_0 \end{aligned} \tag{3.6}$$

has a unique integral solution $w \in C(0, T; X)$. Here we set $v = Q_n u$ and $w = Qu$. Then Q_n and Q are operators from K into $C(0, T; X)$.

LEMMA 3.1. *There exists $V_0 > 0$ such that*

$$QK(V_0) \subset K(V_0).$$

Proof. In [13], Vrabie proved that $QK \subset K$. Therefore it is sufficient to show that there exists $V_0 > 0$ such that $\operatorname{var}(Qu : [0, T]) \leq V_0$ for all $u \in K$ with $\operatorname{var}(u : [0, T]) \leq V_0$. Let $u \in K$ with bounded variation. By (1.6) and (1.7), we have that for each $0 \leq h \leq T$,

$$\|Qu(t+h) - Qu(t)\| \leq \int_0^h \|G(u)(s)\|\, ds + h|Au_0| + \int_0^t \|G(u)(s+h) - G(u)(s)\|\, ds \tag{3.7}$$

(see [3, p. 132]). Then from (3.1) and the condition (C2) we obtain

$$\operatorname{var}(Qu : [0, T]) \leq T\big(M + |Au_0| + k(R)(1 + \operatorname{var}(u : [0, T]))\big), \tag{3.8}$$

where $R = r + \|u_0\|$. Therefore if we set $V_0 = T(M + |Au_0| + k(R))(1 - Tk(R))^{-1}$, it follows that $QK(V_0) \subset K(V_0)$. □

Here we note that the condition (iia) in §2 is equivalent to the following condition [6]:

(iia′) For each $M' > 0$ the set

$$\{x \in D(A); \|x\| \leq M' \text{ and } \|y\| \leq M' \text{ for some } y \in Ax\}$$

is relatively compact in X.

In the following lemma, we use the condition (iia′) instead of (iia).

LEMMA 3.2. *The set $\bigcup_n \{J_n Q_n u(t); 0 \leq t \leq T, u \in K(V_0)\}$ is relatively compact in X.*

Proof. By the same argument as in the proof of Lemma 3.1, we have that for each $u \in K(V_0)$ and $n \geq 1$,

$$\begin{aligned} \|Q_n u(t+h) - Q_n u(t)\| &\leq h\big(M + |Au_0| + k(R)(1 + \operatorname{var}(u: [0,T]))\big) \\ &\leq h\big(M + |Au_0| + k(R)(1+V_0)\big). \end{aligned} \tag{3.9}$$

Since $Q_n u$ is differentiable on $[0,T]$, (3.9) implies

$$\left\| \frac{dQ_n u(t)}{dt} \right\| \leq M + |Au_0| + k(R)(1+V_0) \quad \text{on } [0,T], \tag{3.10}$$

while, by the definition of Q_n,

$$\frac{dQ_n u(t)}{dt} + A_n Q_n u(t) = G(u)(t) \quad \text{on } [0,T]. \tag{3.11}$$

Therefore we have that for every $u \in K(V_0)$ and every $n \geq 1$,

$$\begin{aligned} \|A_n Q_n u(t)\| &\leq \left\| \frac{dQ_n u(t)}{dt} \right\| + \|G(u)(t)\| \\ &\leq 2M + |Au_0| + k(R)(1+V_0) \quad \text{on } [0,T]. \end{aligned} \tag{3.12}$$

Next we show that the set $B = \bigcup_n \{J_n Q_n u(t);\ 0 \leq t \leq T,\ u \in K(V_0)\}$ is bounded. From (1.6), we have

$$\|Q_n u(t) - u_0\| \leq \int_0^t \big(\|G(u)(s)\| + |Au_0|\big)\, ds \leq T\big(M + |Au_0|\big), \tag{3.13}$$

for all $u \in K(V_0)$ and $0 \leq t \leq T$, while by the definition of A_n,

$$\begin{aligned} \|Q_n u(t) - J_n Q_n u(t)\| &\leq n^{-1} \|A_n Q_n u(t)\| \\ &\leq 2M + |Au_0| + k(R)(1+V_0) \quad \text{on } [0,T], \end{aligned} \tag{3.14}$$

for all $u \in K(V_0)$ and $n \geq 1$. From (3.13) and (3.14), we obtain that for each $u \in K(V_0)$, $n \geq 1$ and $0 \leq t \leq T$,

$$\|J_n Q_n u(t)\| \leq \|u_0\| + (2+T)M + (1+T)|Au_0| + k(R)(1+V_0). \tag{3.15}$$

Now we put $M' = \|u_0\| + (2+T)M + (1+T)|Au_0| + k(R)(1+V_0)$. Then since $J_n Q_n u(t) \in D(A)$ and $A_n Q_n u(t) \in A J_n Q_n u(t)$ for all $u \in K(V_0)$ and $0 \leq t \leq T$, we have

$$B \subset \{x \in D(A);\ \|x\| \leq M' \text{ and } \|y\| \leq M' \text{ for some } y \in Ax\}.$$

Therefore by (iia′) we obtain that B is relatively compact. □

Proof of Theorem 2.1. A function $u \in K(V_0)$ is an integral solution of (1.2) if it is a fixed point of Q. As in the proof of [13, Th. 2.1], we show the existence of the fixed points of Q by using Schauder's fixed point theorem. It is easy to see that Q is continuous (see [13]), and $K(V_0)$ is closed convex. To use Schauder's fixed point theorem, we show that $QK(V_0)$ is relatively compact in $C(0,T;X)$. From (3.7), we have that for each $u \in K(V_0)$ and $0 \leq h \leq T$,

$$\|Qu(t+h) - Qu(t)\| \leq h\big(M + |Au_0| + k(R)(1+V_0)\big). \tag{3.16}$$

Therefore $QK(V_0)$ is equicontinuous in $C(0,T;X)$. Then it is sufficient to show that the set $B' = \{Qu(t); u \in K(V_0),\ 0 \leq t \leq T\}$ is relatively compact in X. Fix $u \in K(V_0)$ and $t \in [0,T]$. Since $Q_n u(t)$ converges to $Qu(t)$ (cf. [3, Chapt. III, Lem. 2.1]), and

$$\lim_n \|Q_n u(t) - J_n Q_n u(t)\| = \lim_n n^{-1} \|A_n Q_n u(t)\| = 0,$$

we obtain that $Qu(t)\in\bar{B}=\overline{\cup_n\{J_nQ_nu(t);u\in K(V_0),\,0\le t\le T\}}$. Since $\bar{B}$ is compact in X, the set B' is relatively compact in X. Therefore by Ascoli's theorem, we have that $QK(V_0)$ is relatively compact in $C(0,T;X)$, which completes the proof. □

Proof of Corollary 2.1. Since X is reflexive, (3.9) implies that Q_nu is differentiable almost everywhere on $[0,T]$ for $u\in K(V_0)$. Then by the same argument as in the proof of Theorem 2.1, we obtain the conclusion of Corollary 2.1. □

4. The case where A is weakly closed. In this section, we consider the problem (1.1) under the assumption of weak closedness of A. First we note that each function f: $[0,a]\times U\to X$ $(U\subset X)$ generates a function G: $C(0,a;U)\to C(0,a;X)$.

THEOREM 4.1. *Let X be a reflexive Banach space and $A\subset X\times X$ be a weakly closed m-accretive set. Let $U\subset X$ be an open set and g: $[0,a]\times U\to X$ satisfy the following*:

(C3) *g: $[0,a]\times U\to X$ is a continuous mapping*;

(C4) *$g(t,\cdot)$ is weakly continuous for all $0\le t\le a$, i.e. if $x_\lambda\to x$, then $g(t,x_\lambda)\to g(t,x)$ for all $0\le t\le a$.*

In addition, assume that the G: $C(0,a;U)\to C(0,a;X)$ generated by g satisfies (C2). *Then for each $u_0\in D(A)\cap U$, there exists $T\in(0,a]$ such that* (1.1) *has a strong solution on* $[0,T]$.

COROLLARY 4.1. *Let X,U and g be as in Theorem* 4.1. *Let A be linear and m-accretive. Then for each $u_0\in D(A)\cap U$, there exists $T\in(0,a]$ such that* (1.1) *has a strong solution $u\in C(0,T;U)$ and it satisfies*

$$u(t)=S(t)u_0+\int_0^t S(t-s)f(s,u(s))\,ds,\qquad 0\le t\le T, \tag{4.1}$$

where $\{S(t)\}$ is the semigroup generated by A.

To prove Theorem 4.1, we need the following lemmas.

LEMMA 4.1. *Let X be reflexive. A subset $D\subset C(0,b;X_w)$ is relatively sequentially compact if the set $\{h(t)\colon h\in D,\,0\le t\le b\}$ is bounded and D is equicontinuous in the norm topology.*

Lemma 4.1 is a variant of the Ascoli theorem (see [1, Lem. 4.4]).

LEMMA 4.2. *Let X be reflexive and $V\subset X$ be an open set. Let f: $[0,b]\times V\to X$ be a continuous function such that $f(t,\cdot)$ is weakly continuous for every $0\le t\le b$. Then for each $v_0\in V$, there exists $T\in(0,b]$ such that the initial value problem*

$$\begin{aligned}&\frac{dv(t)}{dt}=f(t,v(t)),\qquad 0\le t\le T,\\ &v(0)=v_0\end{aligned} \tag{4.2}$$

has a strong solution on $[0,T]$.

Proof. Let $v_0\in V$. Since f is continuous, we can choose positive numbers r',M' and T' such that

$$\|f(t,v)\|\le M'\quad\text{for all }(t,v)\in[0,T']\times B(v_0,r'). \tag{4.3}$$

By [11, Lem. 1] we have that for given $\varepsilon>0$ there exists a continuous function f_ε: $[0,T']\times B(r',v_0)\to X$ such that

$$\sup\{\|f(t,v)-f_\varepsilon(t,v)\|;\,(t,v)\in[0,T']\times B(v_0,r')\}<\varepsilon, \tag{4.4}$$

and a strong solution $w\in C(0,T';X)$ of

$$\begin{aligned}&\frac{dw(t)}{dt}=f_\varepsilon(t,w(t)),\qquad 0\le t\le T',\\ &w(0)=v_0\end{aligned} \tag{4.5}$$

exists. (4.4) implies that

$$\left\| w(t) - \int_0^t f(s, w(s))\, ds \right\| \leq \varepsilon T', \qquad 0 \leq t \leq T', \tag{4.6}$$

and

$$\left\| \frac{dw(t)}{dt} \right\| \leq M' + \varepsilon \quad \text{for all } 0 \leq t \leq T'. \tag{4.7}$$

Therefore from Lemma 4.1 we obtain that there exists a sequence $\{u_n\} \subset C(0, T'; B(r', v_0))$ such that $u_n(0) = v_0$ for all $n \geq 1$,

$$\left\| u_n(t) - \int_0^t f(s, u_n(s))\, ds \right\| \leq n^{-1} \tag{4.8}$$

for all $n \geq 1$ and u_n converges in $C(0, T; X_w)$ to a point $u \in C(0, T; X_w)$. Then it is easy to see that the limit point u of $\{u_n\}$ is a strong solution of (4.2). □

LEMMA 4.3. *Let X be reflexive and $A \subset X \times X$ be a weakly closed m-accretive set. Then*:

(a) *For each $n \geq 1$, A_n is weakly continuous, i.e., A_n: $X_w \to X_w$ is continuous.*

(b) *If $\{x_{n_i}\} \subset X$ such that $x_{n_i} \to x$ and $A_{n_i} x_{n_i} \to y$, as $i \to \infty$, then $y \in Ax$.*

Proof. (a) Let $n \geq 1$. It is sufficient to prove that J_n is weakly continuous. Let $x_\lambda \to x$ and put $y_\lambda = J_n x_\lambda$. Then there exists $\{z_\lambda\}$ such that $z_\lambda \in n^{-1} A y_\lambda$ and $x_\lambda = y_\lambda + z_\lambda$. Since J_n is a Lipschitz mapping, $\{y_\lambda\}$ is bounded and so $\{z_\lambda\}$ is bounded. If we choose a subnet $\{y_{\lambda_s}\} \subset \{y_\lambda\}$ and a subnet $\{z_{\lambda_s}\} \subset \{z_\lambda\}$ such that $y_{\lambda_s} \to y$ and $z_{\lambda_s} \to z$, respectively, then from the hypothesis we have that $z \in Ay$. Therefore we obtain that $y_\lambda = J_n x_\lambda \to J_n x = y$. This completes the proof of (a).

(b) Put $y_{n_i} = J_{n_i} x_{n_i}$ and $z_{n_i} = A_{n_i} x_{n_i}$ for $i \geq 1$. Then $z_{n_i} \in A y_{n_i}$ for all $i \geq 1$. Since $\{z_{n_i}\}$ is bounded and

$$x_{n_i} - y_{n_i} = (I - J_{n_i}) x_{n_i} = n_i^{-1} A_{n_i} x_{n_i} = n_i^{-1} z_{n_i} \tag{4.9}$$

for all $i \geq 1$, we have that $y_{n_i} \to x$ and $z_{n_i} \to y$. Then the hypothesis implies $y \in Ax$. □

Proof of Theorem 4.1. Let G be the function generated by g and let M, r, T, V_0, $K, K(V_0)$, Q and Q_n be as in §3. From Lemma 4.2, we obtain that there exists a sequence $\{u_n\} \subset C(0, T; X)$ such that $Q_n u_n = u_n$ for $n \geq 1$. Since each u_n is of bounded variation, by (C2) and [3, (2.19), p. 132] we have that

$$\begin{aligned} \operatorname{var}(u_n; [0, T]) &= \operatorname{var}(Q_n u_n; [0, T]) \\ &\leq T\big(M + |Au_0| + k(R)(1 + \operatorname{var}(u_n; [0, T]))\big). \end{aligned} \tag{4.10}$$

Then it follows that

$$\operatorname{var}(u_n; [0, T]) \leq T\big(M + |Au_0| + k(R)(1 + V_0)\big) \leq V_0. \tag{4.11}$$

Therefore $\{u_n\} \subset K(V_0)$. Then by (3.9), $\{u_n\}$ is equicontinuous in the norm topology. Also from (3.13), we have that the set $\{Q_n u_n(t): 0 \leq t \leq T,\ n \geq 1\}$ is bounded in X. Therefore $\{u_n\}$ is relatively compact in $C(0, T; X_w)$. Let $\{u_{n_i}\}$ be a subsequence of $\{u_n\}$ which converges in $C(0, T; X_w)$ to a point $u \in C(0, T; X_w)$. Hence we set $v(t) = \text{weak-lim} A_{n_i} u_{n_i}(t)$ for all $0 \leq t \leq T$. Then we have

$$\begin{aligned} u(t) &= \text{weak-lim}_i u_{n_i}(t) \\ &= \text{weak-lim} \int_0^t \big(-A_{n_i} u_{n_i}(s) + g(s, u_{n_i}(s))\big)\, ds \\ &= \int_0^t (-v(s) + g(s, u(s)))\, ds. \end{aligned} \tag{4.12}$$

By Lemma 4.3, $v(t)\in Au(t)$ for all $0\le t\le T$. Therefore we obtain that

$$\frac{du(t)}{dt}+Au(t)\ni g(t,u(t)) \quad \text{a.e. on } [0,T]. \tag{4.13}$$

Since $u(0)=\text{weak-lim}\, u_{n_i}(0)=u_0$, u is a strong solution of (1.1). □

Remark. Any linear m-accretive operator is weakly closed since such an operator is a closed operator. In Hilbert spaces, an m-accretive set A is weakly closed if A satisfies the condition (iia).

5. Examples. Throughout this section, Ω is a bounded open subset of R^n with sufficiently smooth boundary Γ. $H^k(\Omega)$ and $H_0^k(\Omega)$ stand for Sobolev spaces on Ω.

Example 1. We consider a nonlinear differential operator of the form

$$Au=\sum_{|\alpha|\le n}(-1)^{|\alpha|}D^\alpha A_\alpha(x,u,\cdots,D^n u), \tag{5.1}$$

where $A_\alpha(x,z)$ are real functions defined on $\Omega\times R$. $A_\alpha(x,z)$ is measurable in x and continuous in z for all α. In addition, we impose on A the following condition:

$$\sum_{|\alpha|\le n}(A_\alpha(x,z)-A_\alpha(x,y))(z_\alpha-y_\alpha)\ge w\Big(\sum_{|\alpha|\le n}|z_\alpha-y_\alpha|^2\Big), \tag{5.2}$$

for $(z,y)\in R^n\times R^n$, where $w>0$.

Now we consider the nonlinear integrodifferential equation [13]:

$$\frac{\partial u}{\partial t}+\sum_{|\alpha|\le n}(-1)^{|\alpha|}D^\alpha A_\alpha(x,u,\cdots,D^n u)+\int_0^t a(t-s)g(s,u(s))\,ds=0, \tag{5.3}$$

with Dirichlet boundary conditions

$$D^\alpha u=0 \quad \text{on } [0,T]\times\Gamma \quad \text{for } |\alpha|\le n-1 \tag{5.4}$$

and initial condition

$$u(0,x)=u_0(x) \quad \text{on } \Omega. \tag{5.5}$$

By using Theorem 2.1, we improve [13, Th. 6.4]:

THEOREM 5.1. *Let $H=L^2(\Omega)$, $V=H_0^1(\Omega)$ and A: $V\to V$ be the nonlinear operator defined above. Let a: $[0,\infty)\to R$ and g: $[0,\infty)R^n\to R$ be continuous functions. In addition, assume that*

$$a'\in L^1_{\text{loc}}(0,\infty), \tag{5.6}$$

$$|g(t,x)|\le b(t)\|u(t)\|+c, \tag{5.7}$$

where $b\in L^1_{\text{loc}}(0,\infty)$ and $c>0$.

Then for each $x_0\in L(\Omega)$, there exists $T>0$ such that (5.3), (5.4) *and* (5.5) *have a strong solution on* $[0,T]$.

Proof. Let A_H be a operator defined by

$$A_H u=Au \quad \text{for } u\in D(A_H)=\{u\in V: Au\in H\}. \tag{5.8}$$

Then A_H is a maximal monotone operator on H [3], and (5.3), (5.4), (5.5) can be rewritten in the form

$$u(0)=u_0, \tag{5.9}$$

where $G(u)(t)=\int_0^t a(t-s)g(s,u(s))\,ds$.

Since A_H is coercive by (5.2) and V is compactly imbedded in H, we can see that $(I+\lambda A)^{-1}$ is compact for $\lambda>0$. It is easy to see that for each $T>0$, G: $C(0,T;H)\to C(0,T;H)$ is continuous. Then in order to apply Theorem 2.1, it is sufficient to show that G satisfies the condition (C2). Now fix $T'>0$. Then for each $0\le t\le T'$,

$$\begin{aligned}\operatorname{var}(G(u)\colon [0,t]) &= \int_0^t \left|\frac{d}{dt}G(u)(s)\right| ds \\ &\le \big(a(0)+\|a'\|_{L^1(0,T')}\big)\|g(u)\|_{L^1(0,t;H)} \\ &\le \big(a(0)+\|a'\|_{L^1(0,T')}\big)\big(\|b\|_{L^1(0,T)}\|u\|_{L^\infty(0,t;H)}+cT'\big).\end{aligned}$$

Therefore (C2) holds, which completes the proof. □

Example 2. Let A be as in Example 1. We consider the following differential equation:

$$\frac{\partial u}{\partial t}+\sum_{|\alpha|\le n}(-1)^{|\alpha|}D^\alpha A_\alpha(x,u,\cdots,D^n u)+\left(\int_\Omega b(x)u(t,x)\,dx\right)u(t,x)=0, \tag{5.10}$$

where $b\in L^2(\Omega)$, with boundary conditions (5.4) and initial value condition (5.5). Then (5.10) and (5.4) can be rewritten in the form:

$$\frac{du}{dt}+A_H u(t)+\langle b,u\rangle u(t)=0,\qquad 0\le t\le T, \tag{5.11}$$

where $\langle\cdot,\cdot\rangle$ stands for the inner product in H.

Then by applying Theorem 2.1, we can see that there exists $T>0$ such that (5.10), (5.4) and (5.5) have a strong solution on $[0,T]$.

Acknowledgment. The author wishes to express his hearty thanks to Professor W. Takahashi, H. Umegaki and the referee for many suggestions and advice in the course of preparing the present paper.

REFERENCES

[1] N. U. AHMED AND K. L. TEO, *Optimal control of systems governed by a class of nonlinear evolution equations in a reflexive Banach space*, J. Optim. Theory Appl., 25 (1978), 59–81.

[2] V. BARBU, *Continuous perturbations of nonlinear m-accretive operators in Banach spaces*, Boll. Un. Mat. Ital., 6 (1972), pp. 270–278.

[3] _____, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Editura Academiei, Bucuresti-Noordhoff, 1976.

[4] P. BENILAN, *Solutions faibles d'équations d'évolution dans un espace relexif*, in Seminaire Deny sur les semigroupes nonlinéaires, Orsay, 1970–1971.

[5] _____, *Solutions integrales d'équations d'évolution dans un espace de Banach*, C. R. Acad. Sci. Paris, 274 (1972), pp. 47–50.

[6] H. BRÉZIS, *New results concerning monotone operators and nonlinear semigroups*, in Lecture Notes of the Research Institute for Mathematical Sciences, Kyoto Univ., 258 (1975), 2–27.

[7] F. E. BROWDER, *Nonlinear equations of evolution*, Ann. Math., 80 (1964), pp. 485–523.

[8] M. G. CRANDALL AND J. NOHEL, *An abstract differential equation and a related Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.

[9] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic Press, New York and London, 1960.

[10] T. KATO, *Nonlinear semigroups and evolution equations*, J. Math. Soc. Japan, 19 (1967), pp. 508–520.

[11] A. LASOTA AND J. A. YORKE, *The generic property of existence of solutions of differential equations in Banach space*, J. Differential Equations, 13 (1973), pp. 1–13.

[12] A. PAZY, *A class of semi-linear equations of evolution*, Israel J. Math., 20 (1975), pp. 23–36.

[13] I. I. VRABIE, *The nonlinear version of Pazy's local existence theorem*, Israel J. Math., 32 (1979), pp. 221–235.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0009 $01.25/0

CONFORMAL KILLING TENSORS AND VARIABLE SEPARATION FOR HAMILTON–JACOBI EQUATIONS*

E. G. KALNINS† AND WILLARD MILLER, JR.‡

Abstract. Every separable coordinate system for the Hamilton–Jacobi equation $g^{ij}W_iW_j=0$ corresponds to a family of $n-1$ conformal Killing tensors in involution, but the converse is false. For general n we find a practical characterization of those families of conformal Killing tensors that correspond to variable separation, orthogonal or not.

1. Introduction. This paper is devoted to the separation of variables problem for the Hamilton–Jacobi equation

$$g^{ij}\partial_{x^i}W\partial W_{x^j}=0, \qquad g^{ij}=g^{ji}, \qquad 1\le i,j\le n \tag{1.1}$$

and the explicit relation between variable separation and second order conformal Killing tensors on the (local) manifold V_n with metric tensor $\{g_{ij}\}$ analytic in the local coordinates $\{x^i\}$. (Here all coordinates and tensors are complex valued and we adopt the notation in Eisenhart's book [1].) Equation (1.1) is intimately related to the separation of variables problem for the Laplace or wave equation,

$$\frac{1}{\sqrt{g}}\partial_{x^i}\left(\sqrt{g}\,g^{ij}\partial_{x^j}\psi\right)=0, \qquad g=\det(g_{ij}). \tag{1.2}$$

It is straightforward to show that any coordinate system yielding (product) R-separation of (1.2) also yields (additive) separation of (1.1). (We have also shown for flat space and $n=3,4$ that the converse holds, i.e., the two equations separate in exactly the same coordinate systems, orthogonal or not [2], [3].)

In 1891 Stäckel [4] showed that (1.1) is additively separable in the orthogonal coordinate system $\{x^i\}$ if and only if there exists a nonzero function $Q(x^j)$ such that the metric $d\hat{s}^2$ where

$$ds^2=g_{ij}dx^i\,dx^j=H_j^2(dx^j)^2=Qh_j^2(dx^j)^2=Q\,d\hat{s}^2 \tag{1.3}$$

can be expressed in *Stäckel form*:

$$h_i^2=\frac{\Theta}{\Theta^{i1}}, \qquad 1=1,\cdots,n \tag{1.4}$$

where Θ is a Stäckel determinant, $\Theta=\det(\theta_{kl})$, $(\theta_{kl}(x^k))$ is a Stäckel matrix (row k depends only on the variable x^k), and Θ^{i1} is the $(i,1)$-cofactor of this matrix. Thus the condition for additive separation of (1.1) in coordinates $\{x^j\}$ is that ds^2 is conformal to a metric $d\hat{s}^2$ in Stäckel form. Separable solutions of (1.1) take the form $W=\sum_{i=1}^n B_i(x^i)$.

Moon and Spencer [5] show that (1.2) admits orthogonal R-separable solutions, i.e., solutions of the form $\psi=e^R\prod_{i=1}^n A_i(x^i)$ where R is a fixed function, if and only if (1) ds^2 is conformal (with factor Q^{-1}) to a Stäckel form metric $d\hat{s}^2$, (2) that

$$\frac{Q\mathcal{H}e^{2R}}{\Theta}=\prod_{i=1}^n f_i(x^i), \qquad \mathcal{H}=H_1H_2,\cdots,H_n, \tag{1.5}$$

*Received by the editors October 31, 1980, and in revised form December 9, 1981.

†Mathematics Department, University of Waikato, Hamilton, New Zealand.

‡School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by the National Science Foundation under grant MCS 78-26216.

and (3) that e^R satisfy

$$\sum_{j=1}^{n} \frac{\Theta^{j1}}{\Theta f_j} \partial_{x^j}\left(f_j \partial_{x^j} e^{-R}\right) + \alpha e^{-R} = 0 \tag{1.6}$$

where α is a constant. (See, however, [6] for a discussion of condition (1.6).) In practice, to determine the orthogonal R-separable coordinate systems for the Laplace equation on the manifold V_n, one first finds all orthogonal separable systems for the Hamilton–Jacobi equation (1.1) and then determines for each system whether or not conditions (1.5) and (1.6) can be satisfied. For nonorthogonal coordinates the relationship is similar but more complicated, see [2], [3].

The relation between variable separation for (1.1) and conformal Killing tensors on V_n is most conveniently presented in terms of the symplectic structure on the cotangent bundle $\tilde{V}_n$ of this manifold. Corresponding to local coordinates $\{x^j\}$ on V_n we introduce coordinates $\{x^j, p_j\}$ on $\tilde{V}_n$. New coordinates $\{\hat{x}^k(x^l)\}$ on V_n correspond to coordinates $\{\hat{x}^k, \hat{p}_k\}$ on $\tilde{V}_n$ where $\hat{p}_k = p_l \partial x^l / \partial \hat{x}^k$. The *Poisson bracket* of two functions $F(x^j, p_j)$, $G(x^j, p_j)$ on $\tilde{V}_n$ is given by

$$[F, G] = \partial_{x^l} F \partial_{p_l} G - \partial_{p_l} F \partial_{x^l} G. \tag{1.7}$$

Let

$$H = g^{ij} p_i p_j. \tag{1.8}$$

A *first order symmetry* of (1.1) is a linear function L in the momenta p_j,

$$L = \xi^j(x) p_j \tag{1.9}$$

such that

$$[L, H] = \rho(x) H \tag{1.10}$$

for some analytic function ρ. Clearly, L is a symmetry if and only if $\{\xi^j\}$ is a conformal Killing vector for V_n [1]. Indeed it is straightforward to show that (1.10) is equivalent to

$$\xi_{i,j} + \xi_{j,i} = \rho g_{ij} \tag{1.11}$$

where $\xi_{i,j}$ is the jth covariant derivative of ξ_i. Similarly a *second order symmetry* of (1.1) is a quadratic function

$$A = a^{ij}(x) p_i p_j, \qquad a^{ij} = a^{ji}, \tag{1.12}$$

such that

$$[A, H] = \left(Q^l(x) p_l\right) H \tag{1.13}$$

where the $\{Q^l\}$ are analytic. Condition (1.13) is equivalent to

$$a_{ij,k} + a_{ki,j} + a_{jk,i} = \frac{1}{2}\left(Q_i g_{jk} + Q_k g_{ij} + Q_j g_{ki}\right), \tag{1.14}$$

i.e., $\{a^{ij}\}$ (or $\{a_{ij}\}$) is a *conformal Killing tensor of order* 2. It is obvious that $\rho(x)H$ is a (trivial) conformal Killing tensor for any analytic function ρ. Thus by addition of multiples ρH of H if necessary, one could assume that every nontrivial conformal Killing tensor is traceless, $a^l_l = 0$. (We shall ordinarily not make this assumption.) Note that then the Q_i can be expressed simply in terms of the components of a traceless A: $Q_i = (4/(n+2))a^l_{1,l}$. For future use we also note that the condition for two quadratic functions A and $B = b^{ij} p_i p_j$ to be in involution, i.e., $[A, B] = 0$, is

$$a_{ij,l} b^l_k + a_{ki,l} b^l_j + a_{jk,l} b^l_i = b_{ij,l} a^l_k + b_{ki,l} a^l_j + b_{jk,l} a^l_i. \tag{1.15}$$

We can now state the basic relation between separation of variables for (1.1) and conformal Killing tensors: To every orthogonal coordinate system $\{y^i\}$ which permits additive separation of variables in (1.1), there correspond $n-1$ second order conformal Killing tensors $A_1,\cdots,A_{n-1}$ which are in involution and such that $\{H,A_1,\cdots,A_{n-1}\}$ is linearly independent. The separable solutions $W=\Sigma_{k=1}^n W^{(k)}(y^k)$ are characterized by the relations

$$(1.16)\qquad H(y^j,p_j)=0,\qquad A_l(y^j,p_j)=\lambda_l,\qquad l=1,\cdots,n-1,\quad p_j=\partial_{y^j}W,$$

where $\lambda_1,\cdots,\lambda_{n-1}$ are the separation constants. The basis tensors A_l are of course not unique, but the space spanned by these tensors is uniquely determined. (A new proof of this correspondence is contained in Theorems 4 and 7 to follow. Expressions (1.16) are then obvious from Stäckel's construction.) For nonorthogonal separable coordinates, the same characterization is valid except that one or more of the A_l are conformal Killing vectors. For $n\le 4$ all possible separable systems and their corresponding conformal Killing tensors have been explicitly determined [2], [3].

A remaining problem with the theory is that there exist involutive families of $n-1$ conformal Killing tensors that are not related to any separable coordinate system. In this paper we give a complete solution to this problem. That is, we provide directly verifiable necessary and sufficient conditions for a family of conformal Killing tensors to determine a separable coordinate system for (1.1), and we show how to compute the separable coordinates from the given tensors. In §2 we study the case of orthogonal separable coordinates where the Killing tensor characterization is especially simple. Finally, in §3 we treat the general nonorthogonal case. The results of this paper are a nontrivial extension of the results in [7], [8] for the equation $g^{ij}\partial_i W\partial_j W=E$ with $E\neq 0$.

2. The orthogonal case. Let $\{x^j\}$ be a local orthogonal coordinate system on V_n and let $ds^2=g_{ij}dx^i dx^j=H_j^2(dx^j)^2$ be the metric for V_n as expressed in these coordinates. It follows from (1.3) and (1.4) that the Hamilton–Jacobi equation (1.1) is separable in the $\{x^i\}$ if and only if there exists an analytic function $Q(x)$ such that $H_j=Qh_j$ where the metric $d\hat{s}^2=h_j^2(dx^j)^2$ is in Stäckel form. We begin our study of such "conformally Stäckel" metrics by deriving a more convenient characterization for them.

It is well known that the metric $d\hat{s}^2$ is in Stäckel form with respect to the coordinates $\{x^j\}$ if and only if the conditions

$$(2.1)\qquad \frac{\partial^2\ln h_i^2}{\partial x^j\partial x^k}-\frac{\partial\ln h_i^2}{\partial x^j}\frac{\partial\ln h_i^2}{\partial x^k}+\frac{\partial\ln h_i^2}{\partial x^j}\frac{\partial\ln h_j^2}{\partial x^k}+\frac{\partial\ln h_i^2}{\partial x^k}\frac{\partial\ln h_k^2}{\partial x^j}=0,\qquad j\neq k,$$

are satisfied [1, App. 13]. Let $d\tilde{s}^2=K_j^2(dx^j)^2$ where $K_j^2=h_j^2/h_n^2$; in particular $K_n^2=1$. A straightforward computation using (2.1) yields

LEMMA 1. *If the metric $d\hat{s}^2=h_j^2(dx^j)^2$ is in Stäckel form then so is the metric $d\tilde{s}^2=h_n^{-2}d\hat{s}^2$.*

Now let $ds^2=H_j^2(dx^j)^2=Q\,d\hat{s}^2$. If $d\hat{s}^2$ is in Stäckel form, then by Lemma 1 the metric $H_n^{-2}ds^2=h_n^{-2}d\hat{s}^2$ is also in Stäckel form. Conversely, if $H_n^{-2}ds^2$ is in Stäckel form then $ds^2=H_n^2(H_n^{-2}ds^2)$ is conformal to a Stäckel form metric. This proves

LEMMA 2. *$ds^2=H_j^2(dx^j)^2$ is conformal to a Stäckel form metric if and only if the coefficients H_j^2 satisfy the conditions*

$$(2.2)\qquad \frac{\partial^2\ln K_i^2}{\partial x^j\partial x^k}-\frac{\partial\ln K_i^2}{\partial x^j}\frac{\partial\ln K_i^2}{\partial x^k}+\frac{\partial\ln K_i^2}{\partial x^j}\frac{\partial\ln K_j^2}{\partial x^k}+\frac{\partial\ln K_i^2}{\partial x^k}\frac{\partial\ln K_k^2}{\partial x^j}=0,\qquad j\neq k,$$

where $K_j^2=H_j^2/H_n^2$.

Note that for $i=n$, equations (2.2) are satisfied identically and for $k=n$ they read

$$\frac{\partial^2 \ln K_i^2}{\partial x^j \partial x^n}+\frac{\partial \ln K_i^2}{\partial x^j}\frac{\partial \ln\left(K_j^2/K_i^2\right)}{\partial x^n}=0, \qquad j\neq n. \tag{2.3}$$

THEOREM 1. *Let A be a second order conformal Killing tensor such that the n roots $\rho_1(x),\cdots,\rho_n(x)$ of the characteristic equation*

$$\det(a_{ij}-\rho g_{ij})=0 \tag{2.4}$$

are pairwise distinct. Furthermore, suppose the eigenvector fields corresponding to these n roots are normalizable, i.e., there exists a coordinate system $\{y^j\}$ on V_n such that

$$ds^2=g_{ij}dx^i dx^j=H_j^2\left(dy^j\right)^2, \qquad \psi=a_{ij}dx^i dx^j=\rho_j H_j^2\left(dy^j\right)^2. \tag{2.5}$$

Then the Hamilton–Jacobi equation (1.1) *is separable in the coordinates $\{y^j\}$.*

Proof. Conditions (1.14) for A are equivalent to

$$\partial_{y^l}\ln\left(\frac{\rho_l-\rho_k}{H_k^2}\right)=0, \qquad l\neq k. \tag{2.6}$$

Setting $\mu_\alpha=\rho_\alpha-\rho_n$, $\alpha=1,\cdots,n-1$, we see that these equations can be written in the form

$$\text{a)}\quad \partial_{y^\alpha}\ln\left(\frac{\mu_\alpha}{H_n^2}\right)=0, \qquad \alpha=1,\cdots,n-1,$$

$$\text{b)}\quad \partial_{y^n}\ln\left(\frac{\mu_\alpha}{H_\alpha^2}\right)=0, \tag{2.7}$$

$$\text{c)}\quad \partial_{y^\alpha}\ln\left(\frac{\mu_\alpha-\mu_\beta}{H_\beta^2}\right)=0, \qquad 1\leq\alpha,\beta\leq n-1, \quad \alpha\neq\beta,$$

or

$$\begin{aligned}
&\partial_\alpha\mu_\beta=(\mu_\beta-\mu_\alpha)\partial_\alpha\ln\left(H_\beta^2\right)+\mu_\alpha\partial_\alpha\ln H_n^2, \qquad \alpha\neq\beta,\\
&\partial_\alpha\mu_\alpha=\mu_\alpha\partial_\alpha\ln\left(H_n^2\right),\\
&\partial_n\mu_\alpha=\mu_\alpha\partial_n\ln\left(H_\alpha^2\right).
\end{aligned} \tag{2.8}$$

The integrability conditions $\partial_i\partial_j\mu_\alpha=\partial_j\partial_i\mu_\alpha$ for the system (2.8) can be written in the form

$$\begin{aligned}
&(\mu_\rho-\mu_\alpha)\left[\partial_{\alpha\beta}\ln\left(\frac{H_\beta^2}{H_n^2}\right)+\partial_\alpha\ln\left(\frac{H_\beta^2}{H_n^2}\right)\partial_\beta\ln\left(\frac{H_\alpha^2}{H_n^2}\right)\right]=0, \qquad \alpha\neq\beta,\\
&(\mu_\gamma-\mu_\alpha)\left[\partial_{\alpha\gamma}\ln\left(\frac{H_\beta^2}{H_n^2}\right)-\partial_\alpha\ln\left(\frac{H_\beta^2}{H_n^2}\right)\partial_\gamma\ln\left(\frac{H_\beta^2}{H_n^2}\right)\right.\\
&\qquad\left.+\partial_\alpha\ln\left(\frac{H_\beta^2}{H_n^2}\right)\partial_\gamma\ln\left(\frac{H_\alpha^2}{H_n^2}\right)+\partial_\alpha\ln\left(\frac{H_\gamma^2}{H_n^2}\right)\partial_\gamma\ln\left(\frac{H_\beta^2}{H_n^2}\right)\right]=0,\\
&\qquad\qquad \alpha,\beta,\gamma \text{ pairwise distinct},\\
&(\mu_\beta-\mu_\alpha)\left[\partial_{\alpha n}\ln\left(\frac{H_\beta^2}{H_n^2}\right)-\partial_\alpha\ln\left(\frac{H_\beta^2}{H_n^2}\right)\partial_n\left(\frac{H_\beta^2}{H_n^2}\right)\right.\\
&\qquad\left.+\partial_\alpha\ln\left(\frac{H_\beta^2}{H_n^2}\right)\partial_n\ln\left(\frac{H_\alpha^2}{H_n^2}\right)\right]=0, \qquad \alpha\neq\beta.
\end{aligned} \tag{2.9}$$

Since the ρ_i are pairwise distinct by assumption, we have $\mu_\beta - \mu_\alpha \neq 0$ for $\alpha \neq \beta$, so conditions (2.9) become

$$\frac{\partial^2 \ln K_i^2}{\partial y^j \partial y^k} - \frac{\partial \ln K_i^2}{\partial y^j}\frac{\partial \ln K_i^2}{\partial y^k} + \frac{\partial \ln K_i^2}{\partial y^j}\frac{\partial \ln K_j^2}{\partial y^k} + \frac{\partial \ln K_i^2}{\partial y^k}\frac{\partial \ln K_k^2}{\partial y^j} = 0, \qquad j \neq k, \tag{2.10}$$

where $K_i^2 = H_i^2/H_n^2$, $i = 1, \cdots, n$. It follows from Lemma 2 that $H_j^2(dy^j)^2$ is conformal to a Stäckel metric, hence (1.1) separates in the coordinates $\{y^j\}$. Q.E.D.

Note that if (1.1) is separable in the coordinates $\{y^j\}$, then equations (2.10) hold and the integrability conditions for the system (2.8) are satisfied identically. Thus (2.8) admits a basis of $n-1$ vector solutions $\{\rho_j^{(\beta)}\}$, $\beta = 1, \cdots, n-1$. This proves

THEOREM 2. *Necessary and sufficient conditions that the metric* $ds^2 = g_{ij}dx^i dx^j = H_j^2(dy^j)^2$ *on* V_n *is conformal to a Stäckel form metric with respect to the coordinates* $\{y^j\}$ *are*:

1) *The space admits* $n-1$ *conformal Killing tensors* $a_{ij}^{(\beta)}$, $\beta = 1, \cdots, n-1$ *such that the n tensors* $\{g_{ij}, a_{ij}^{(\beta)}\}$ *form a linearly independent set at each point* **x**.

2) *The roots* $\rho^{(\beta)}$ *for each of the characteristic equations* $\det(a_{ij}^{(\beta)} - \rho^{(\beta)} g_{ij}) = 0$ *are simple.*

3)

$$\left(a_{ij}^{(\beta)} - \rho_h^{(\beta)} g_{ij}\right)\lambda_{(h)}^i = 0, \qquad h = 1, \cdots, n, \quad \beta = 1, \cdots, n-1, \tag{2.11}$$

where $\rho_1^{(\beta)}, \cdots, \rho_n^{(\beta)}$ *are the roots of* $a_{ij}^{(\beta)}$ *and* $\lambda_{(h)}^i = \partial x^i / \partial y^h$.

Note that condition 3) requires the vector fields $\lambda_{(1)}^i, \cdots, \lambda_{(n)}^i$ to be normal and to satisfy equations (2.11) for all β. Theorem 2 and its proof are patterned after the corresponding theorem due to Eisenhart which relates Killing tensors and (true) Stäckel forms [1], [9]. The theorem is not very useful in a practical sense because of the difficulty in deciding when the vector fields $\{\lambda_{(h)}^i\}$ defined by (2.11) are normalizable, i.e., when there exists an orthogonal coordinate system $\{y^i\}$ such that $\{\lambda_{(h)}^i\}$ is orthogonal to the coordinate surface $y^h = \text{const}$, for each $h = 1, \cdots, n$.

To solve this problem we recall some classical results in differential geometry that can be found in Eisenhart's book [1]. Given a family of orthogonal vector fields $\{\lambda_{(h)}^i(x), 1 \leq h \leq n\}$ we define their *coefficients of rotation* γ_{lhk} by

$$\gamma_{lhk} = \lambda_{(l)i,j}\lambda_{(h)}^i \lambda_{(k)}^j, \qquad 1 < l, h, k \leq n; \tag{2.12}$$

see [1, p. 97]. A necessary and sufficient condition that there exist coordinates $\{y^h\}$ and nonzero invariant functions f_h such that $\lambda_{(h)}^i = (\partial x^i / \partial y^h) f_h$, $h = 1, \cdots, n$, is

$$\gamma_{lhk} = 0, \qquad 1 \leq l, h, k \leq n, \quad h, k, l \text{ pairwise distinct.} \tag{2.13}$$

Let a_{ij} be a tensor field with n roots $\rho_1, \cdots, \rho_n$ (not necessarily distinct) and let $\{\lambda_{(h)}^i\}$ be a corresponding orthonormal set of eigenvectors:

$$(a_{ij} - \rho_h g_{ij})\lambda_{(h)}^i = 0, \qquad h = 1, \cdots, n, \tag{2.14}$$

$$\lambda_{(h)}^i \lambda_{(k)i} = \delta_{hk}, \qquad 1 \leq h, k \leq n. \tag{2.15}$$

It follows easily from (2.12), (2.14) and (2.15) that

$$a_{ij,k}\lambda_{(h)}^i \lambda_{(l)}^j \lambda_{(m)}^k = (\rho_h - \rho_l)\gamma_{hlm}, \qquad h \neq l. \tag{2.16}$$

From (2.13) we find

THEOREM 3 (Eisenhart [1, p. 118]). *If* a_{ij} *has pairwise distinct roots* $\rho_1, \cdots, \rho_n$ *then the vector fields* $\{\lambda_{(h)}^i\}$ *are normalizable if and only if*

$$a_{ij,k}\lambda_{(h)}^i \lambda_{(l)}^j \lambda_{(m)}^k = 0, \qquad i \leq h, l, m \leq n, \quad h, l, m \text{ distinct.} \tag{2.17}$$

This leads us to our fundamental result:

THEOREM 4. *Necessary and sufficient conditions that the orthogonal coordinate system* $\{y^j\}$ *be separable for the Hamilton–Jacobi equation* (1.1) *are the existence of* $n-1$ *quadratic functions* $A^{(\beta)}$, $\beta=1,\cdots,n-1$, (1.12), *such that*:

1) *The* $\{A^{(\beta)}\}$ *are second order symmetries of* (1.1), *i.e., the* $\{a_{ij}^{(\beta)}\}$ *are conformal Killing tensors.*

2) *The* $\{A^{(\beta)}\}$ *are in involution*: $[A^{(\alpha)},A^{(\beta)}]=0$, $1\le\alpha,\beta\le n-1$.

3) *The set* $\{H,A^{(1)},\cdots,A^{(n-1)}\}$ *is linearly independent* (*as n quadratic forms at each point* **x**).

4) *At least one of the quadratic forms, say* $A^{(1)}$, *has pairwise distinct roots.*

5) *In any local coordinate system* $\{x^j\}$ *the quadratic forms satisfy the algebraic commutation property*

$$a_{ij}^{(\alpha)}a_k^{(\beta)j}=a_{ij}^{(\beta)}a_k^{(\alpha)j}. \tag{2.18}$$

(*This property is independent of local coordinates.*)

Proof. We suppose that conditions 1)–5) are satisfied. Conditions 4) and 5) imply that the quadratic forms can be simultaneously diagonalized by a family of orthonormal vector fields. In local coordinates $\{x^j\}$ we have

$$\left(a_{ij}^{(\beta)}-\rho_h^{(\beta)}g_{ij}\right)\lambda_{(h)}^i=0, \qquad h=1,\cdots,n, \quad \beta=1,\cdots,n-1, \tag{2.19}$$

where $\rho_1^{(\beta)},\cdots,\rho_n^{(\beta)}$ are the roots of $a_{ij}^{(\beta)}$ and $\lambda_{(h)}^i\lambda_{(k)i}=\delta_{hk}$. Setting $\rho_h^{(n)}=1$, for $h=1,\cdots,n$ we can express condition 3) as

$$\det\left(\rho_m^{(l)}\right)\neq 0. \tag{2.20}$$

Furthermore, by (1.14), (1.15), (2.16), and (2.19), conditions 1) and 2) imply

$$\det\begin{pmatrix}\rho_l^{(\alpha)} & \rho_h^{(\alpha)} & \rho_m^{(\alpha)}\\ 1 & 1 & 1\\ \gamma_{mhl} & \gamma_{lmh} & \gamma_{hlm}\end{pmatrix}=0, \qquad 1\le\alpha\le n-1, \quad h,l,m \text{ distinct}, \tag{2.21}$$

and

$$\det\begin{pmatrix}\rho_l^{(\alpha)} & \rho_h^{(\alpha)} & \rho_m^{(\alpha)}\\ \rho_l^{(\beta)} & \rho_h^{(\beta)} & \rho_m^{(\beta)}\\ \gamma_{hlm}+\gamma_{lmh} & \gamma_{hlm}+\gamma_{mhl} & \gamma_{mhl}+\gamma_{lmh}\end{pmatrix}=0, \qquad 1\le\alpha<\beta\le n-1. \tag{2.22}$$

From (2.20) and (2.21) we have $\gamma_{mhl}=\gamma_{lmh}=\gamma_{hlm}$. Substituting this result into (2.22) and using (2.20) we find $\gamma_{mhl}=\gamma_{lmh}=\gamma_{hlm}=0$. Thus, by (2.13) the vector fields $\{\lambda_{(h)}^j\}$ are normalizable. It then follows from Theorem 2 that the $\{A^{(\beta)}\}$ determine an orthogonal separable coordinate system $\{y^j\}$.

Conversely, given an orthogonal separable coordinate system $\{y^j\}$ for (1.1), we see from the definition of separability, (e.g., (3.5)), that $H=fH'$ for some function f where H' is in Stäckel form with respect to these coordinates. It follows from [7, Thm. 6], that there exist Killing tensors (with respect to H') $A_1,\cdots,A_{n-1}$ that satisfy properties 2)–5). It is obvious that the A_j are conformal Killing tensors for H. Q.E.D.

3. The general case. We now examine the separation for variables problem for (1.1) for the more general case in which the separable coordinates may be nonorthogonal. Our definition of variable separation is identical with that presented in [2], [3] and is based on a division of the separable coordinates into three classes: *ignorable*, *essential of type* 1 and *essential of type* 2. Let $\{x^j\}$ be a coordinate system on V_n with contravariant metric tensor (g^{ij}) and such that the first n_1 coordinates x^a are essential of

type 1, the next n_2 coordinates x^r are essential of type 2, and the last n_3 coordinates x^α are ignorable, $n=n_1+n_2+n_3$. (In the following, indices a,b,c range from 1 to n, indices r,s,t range from n_1+1 to n_1+n_2, indices α,β,γ range from n_1+n_2+1 to n, and indices i,j,k range from 1 to n.) This means that in terms of the coordinates $\{x^j\}$ the metric satisfies $g^{ik}=Q\hat{g}^{ik}$ where $\partial_\alpha \hat{g}^{ik}=0$, $\alpha=n_1+n_2+1,\cdots,n$, and that the separation equations take the form

$$W_a^2+\sum_{\alpha,\beta} A_a^{\alpha,\beta}(x^a)W_\alpha W_\beta=\Phi_a(x^a,\boldsymbol{\lambda}), \tag{3.1}$$

$$2\sum_\alpha B_r^\alpha(x^r)W_rW_\alpha+\sum_{\alpha,\beta} C_r^{\alpha,\beta}(x^r)W_\alpha W_\beta=\Phi_r(x^r,\boldsymbol{\lambda}), \tag{3.2}$$

$$W_\alpha=\lambda_\alpha. \tag{3.3}$$

Here $A_a^{\alpha,\beta}(=A_a^{\beta,\alpha})$, $C_r^{\alpha,\beta}(=C_r^{\beta,\alpha})$ and Φ_i are defined and analytic in a neighborhood $N\subset C^{n_1+n_2}$ of some given point $(x_0^1,\cdots,x_0^{n_1+n_2})$. Furthermore,

$$\Phi_i(x^i,\boldsymbol{\lambda})=\sum_{j=2}^{n_1+n_2}\lambda_j\theta_{ij}(x^i),\qquad i=1,\cdots,n_1+n_2, \tag{3.4}$$

where the complex parameters $\lambda_1,\cdots,\lambda_n$ are arbitrary and the vectors $\partial_{\lambda_j}\tilde{\Phi}$, $j=2,\cdots,n_1+n_2$ are linearly independent for $\mathbf{x}\in N$.

We say that the coordinates $\{x^j\}$ are *separable* for the H–J equation

$$\sum g^{ij}\partial_i W\partial_j W=0 \tag{3.5}$$

if there exist analytic functions A,B,C,Φ above and functions $U_a(x^i)$, $V_r(x^i)$, analytic in N, such that (3.5) can be written in the form

$$\sum_a U_a\Phi_a+\sum_r V_r\Phi_r=0 \tag{3.6}$$

(identically in the parameters $\lambda_2,\cdots,\lambda_{n_1+n_2}$), where $W=\sum_{j=1}^n W^{(j)}(x^j)$, $W_i=\partial_i W=\partial_i W^{(i)}$.

The functions U_a, V_r are uniquely determined by (3.6) up to an arbitrary multiplicative factor $Q(\mathbf{x})$. To analyse the structure of these solutions it is convenient to introduce an $(n_1+n_2)\times(n_1+n_2)$ Stäckel matrix $(\theta_{ij}(x^i))$, $i,j=1,\cdots,n_1+n_2$ whose first column (not unique) is subject only to the condition $\Theta=\det(\theta_{ij})\neq 0$ and whose remaining columns are determined by (3.4). Then

$$U_a=\frac{Q\Theta^{a1}}{\Theta},\qquad V_r=\frac{Q\Theta^{r1}}{\Theta} \tag{3.7}$$

where Θ^{lm} is the (lm)-cofactor of the matrix (θ_{ij}). The nonzero components of the contravariant metric tensor are thus

$$\begin{aligned} &g^{ab}=\left(\frac{Q\Theta^{a1}}{\Theta}\right)\delta^{ab},\qquad g^{r\alpha}=g^{\alpha r}=\left(\frac{Q\Theta^{r1}}{\Theta}\right)B_r^\alpha(x^r),\\ &\frac{1}{2}g^{\alpha\beta}=Q\left(\sum_a A_a^{\alpha,\beta}(x^a)\frac{\Theta^{a1}}{\Theta}+\sum_r C_r^{\alpha,\beta}(x^r)\frac{\Theta^{r1}}{\Theta}\right),\qquad \alpha\neq\beta,\\ &g^{\alpha\alpha}=Q\left(\sum_a A_a^{\alpha,\alpha}(x^a)\frac{\Theta^{a1}}{\Theta}+\sum_r C_r^{\alpha,\alpha}(x^r)\frac{\Theta^{r1}}{\Theta}\right). \end{aligned} \tag{3.8}$$

Furthermore,

$$\sum_{l=1}^{n_1+n_2} \frac{\Theta^{lm}}{\Theta}\Phi_l = \begin{cases} 0 & \text{if } m=1, \\ \lambda_m & \text{otherwise,} \end{cases} \tag{3.9}$$

so,

$$\begin{aligned} &H(\mathbf{x},\mathbf{p}) \equiv g^{ij}p_i p_j = 0, \\ &A_m(\mathbf{x},\mathbf{p}) \equiv a^{ij}_{(m)} p_i p_j = \lambda_m, \qquad m=2,\cdots,n_1+n_2, \\ &L_\alpha(\mathbf{x},\mathbf{p}) \equiv p_\alpha = \lambda_\alpha, \qquad p_i = \partial_{x^i} W \end{aligned} \tag{3.10}$$

where the nonzero terms of the symmetric quadratic form $(a^{ij}_{(m)})$ are

$$\begin{aligned} &a^{ab}_{(m)} = \left(\frac{\Theta^{am}}{\Theta}\right)\delta^{ab}, \quad a^{r\alpha}_{(m)} = \left(\frac{\Theta^{rm}}{\Theta}\right)B^\alpha_r, \\ &\frac{1}{2}a^{\alpha\beta}_{(m)} = \sum_c A^{\alpha,\beta}_c \frac{\Theta^{cm}}{\Theta} + \sum_r C^{\alpha,\beta}_r \frac{\Theta^{rm}}{\Theta}, \qquad \alpha \neq \beta, \\ &a^{\alpha\alpha}_{(m)} = \sum_c A^{\alpha,\alpha}_c \frac{\Theta^{cm}}{\Theta} + \sum_r C^{\alpha,\alpha}_r \frac{\Theta^{rm}}{\Theta}. \end{aligned} \tag{3.11}$$

It follows immediately from [8, Thm. 2] that

$$\begin{aligned} &\text{(a)} \qquad A_m, L_\alpha \text{ are conformal Killing tensors,} \\ &\text{(b)} \qquad [A_m A_l]=0, \quad [A_m, L_\alpha]=0, \quad [L_\alpha, L_\beta]=0. \end{aligned} \tag{3.12}$$

Note that while relations (3.6) determine the coordinates and the metric in an essentially unique manner, there is some freedom of choice for the conformal Killing tensors A_m, due to the nonuniqueness of the first column in the Stäckel matrix. (This freedom is due to the fact that we may replace A_m by $A_m + f(\mathbf{x})H$ without altering relations (3.10).)

We shall now analyse the structure of these separation equations and their relationship to the commutation properties (3.12). First we derive practical, necessary and sufficient conditions to determine if a given coordinate system $\{x^j\}$ yields separation for the Hamilton–Jacobi equation (1.1). Let g^{ij} be the components of the contravariant metric tensor in these coordinates. It is convenient to reorder the coordinates in a standard form. Let n_3 be the number of ignorable variables x^α. Of the remaining $n-n_3$ variables, suppose n_2 variables x^r have the property $g^{rr}=0$ and the remaining n_1 variables x^a satisfy $g^{aa}\neq 0$. We relable the variables so that $1 \leq a \leq n_1$, $n_1+1 \leq n \leq n_1+n_2$, and $n_1+n_2+1 \leq \alpha \leq n$.

THEOREM 5. *Suppose* (g^{ij}) *is in standard form with respect to the variables* $\{x^i\}$. *The Hamilton–Jacobi equation* (1.1) *is separable for this system if and only if*:

1) *The contravariant metric assumes the form*

$$(g^{ij}) = \begin{array}{c} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \begin{pmatrix} \delta^{ab}H_a^{-2} & 0 & 0 \\ 0 & 0 & H_r^{-2}B_r^\alpha \\ 0 & H_r^{-2}B_r^\alpha & g^{\alpha\beta} \end{pmatrix} \end{array} \begin{matrix} n_1 \\ n_2 \\ n_3 \end{matrix}$$

where $B^\alpha_r = B^\alpha_r(x^r)$.

2) *The metric* $d\tilde{s}^2=\sum_{a=1}^{n_1}H_a^2(dx^a)^2+\sum_{r=n_1+1}^{n_1+n_2}H_r^2(dx^r)^2$ *is conformal to a Stäckel form metric, i.e., relations* (2.2) *hold for* $K_j^2=H_j^2/H_{n_1+n_2}^2$.

3) *For each* $g^{\alpha\beta}(\mathbf{x})$, $g^{\alpha\beta}H_{n_1+n_2}^2$ *is a Stäckel multiplier for the metric* $d\tilde{s}^2/H_{n_1+n_2}^2$, *i.e.*,

$$\partial_{ij}g^{\alpha\beta}+\partial_i g^{\alpha\beta}\partial_j\ln H_i^2+\partial_j g^{\alpha\beta}\partial_i\ln H_j^2+g^{\alpha\beta}\left(\partial_{ij}\ln H_j^2+\partial_i\ln H_j^2\,\partial_j\ln H_i^2\right)=0.$$

Proof. This result follows directly from [8, Thm. 1] and Lemma 2.

THEOREM 6. *Let* (g^{ij}) *be the metric tensor on* V_n *in the coordinates* $\{x^i\}$. *If the Hamilton–Jacobi equation* (1.1) *is separable in these coordinates then there exist a function* $Q(\mathbf{x})$ *and a* κ*-dimensional vector space* $\mathcal{Q}$ *of second order conformal Killing tensors on* V_n *such that*:

1) *Each* L_α *and* $A\in\mathcal{Q}$ *is a (true) Killing tensor for the Hamiltonian* $\hat{H}$, *where* $H=Q(\mathbf{x})\hat{H}$, *and* $\hat{H}\in\mathcal{Q}$.

2) $[A,B]=0$, $[L_\alpha,L_\beta]=0$, $[L_\alpha,A]=0$ *for all* $A,B\in\mathcal{Q}$.

3) *For each of the* n_1 *essential coordinates of type* 1, x^a, *the form* dx^a *is a simultaneous eigenform for each* $A\in\mathcal{Q}$, *with simple root* ρ_a^A.

4) *For each of the* n_2 *essential coordinates of type* 2, x^r, *the form* dx^r *is a simultaneous eigenform for every* $A\in\mathcal{Q}$, *with root* ρ_r^A *of multiplicity* 2. *The root* ρ^A *corresponds to only one eigenform.*

5) $\partial_i(a^{\alpha\beta}-\rho_i^A g^{\alpha\beta})=0$, $i=1,\cdots,n_1+n_2$, $A\in\mathcal{Q}$.

6) $g^{ab}=0$ *if* $a\neq b$; $g^{ar}=g^{a\alpha}=g^{rs}=0$.

7) $\kappa=n+n_3(n_3-1)/2$.

These results are readily obtained from the following theorem. Let $\{x^i\}$ be a local coordinate system for V_n with coordinates divided into three classes containing n_1, n_2 and n_3 variables, respectively. (We call these variables essential of types 1 and 2 or ignorable, respectively, even though they may have nothing to do with variable separation.) Let $H=g^{ij}p_ip_j$.

THEOREM 7. *Suppose there exists a* κ*-dimensional space* $\mathcal{Q}$ *of second order conformal Killing tensors and an* n_3*-dimensional space of Killing vectors with basis* $L_\alpha=p_\alpha$, $\alpha=n_1+n_2+1,\cdots,n$. *Furthermore, suppose conditions* 2)–7) *of Theorem* 6 *are satisfied. Then the Hamilton–Jacobi equation* (1.1) *is separable in the coordinates* $\{x^i\}$. *There exists a Stäckel matrix* $(\theta_{ij}(x^i))$ *such that the Killing tensors* A_1, A_m, $m=2,\cdots,n_1+n_2$, (3.10) *and* $L_\alpha L_\beta=p_\alpha p_\beta$, $n_1+n_2+1\le\alpha\le\beta\le n$, *form a basis for* $\mathcal{Q}$.

Proof. Most of the proof follows closely that of [8, Thm. 3], with the added complication that the elements of $\mathcal{Q}$ are conformal, rather than true, Killing tensors. Conditions 3), 4) and 6) imply that for any $A\in\mathcal{Q}$ we have

$$(a^{ij})=\begin{array}{c} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \begin{pmatrix} \delta^{ab}\rho_a H_a^{-2} & 0 & 0 \\ 0 & 0 & \rho_r g^{r\alpha} \\ 0 & \rho_r g^{\alpha r} & a^{\alpha\beta} \end{pmatrix} \end{array}\begin{matrix} n_1 \\ n_2 \\ n_3 \end{matrix}. \tag{3.13}$$

If $(\boldsymbol{\rho}^A)=(\boldsymbol{\rho}^B)$ for $A,B\in\mathcal{Q}$ it follows from (3.13) and condition 5) that $A-B$ is a linear combination of the $n_3(n_3+1)/2$ conformal Killing tensors $L_\alpha L_\beta=p_\alpha p_\beta$, $\alpha\le\beta$.

The condition (1.13) can be written as

$$\begin{aligned} a^{ij}\partial_j g^{kl}+a^{lj}\partial_j g^{ik}+a^{kj}\partial_j g^{li}-g^{ij}\partial_j a^{kl}-g^{lj}\partial_j a^{ik}-g^{kj}\partial_j a^{li} \\ =Q^i g^{kl}+Q^l g^{ik}+Q^k g^{li}. \end{aligned} \tag{3.14}$$

Setting $(i,k,l)=(a,b,b)$ in (3.14) we obtain

$$\partial_a(\rho_b-\rho_a)=(\rho_b-\rho_a)\partial_a \ln H_b^{-2}, \qquad \partial_a\rho_a=-Q^a H_a^2. \tag{3.15}$$

Setting $(i,k,l)=(a,r,\alpha)$ we find

$$\partial_a(\rho_r-\rho_a)=(\rho_r-\rho_a)\partial_a \ln g^{r\alpha} \quad \text{if } g^{r\alpha}\not\equiv 0 \tag{3.16}$$

and for $(i,k,l)=(a,a,\alpha)$,

$$\partial_r\rho_a+(\rho_a-\rho_r)\partial_r \ln H_a^{-2}+g_{r\beta}Q^\beta, \qquad g^{\alpha s}g_{s\beta}Q^\beta=Q^\alpha. \tag{3.17}$$

The case $(i,k,l)=(a,a,r)$ leads to $Q^r=0$ and $(i,k,l)=(r,\alpha,\beta)$ leads to

$$\begin{aligned}(g^{\beta s}g^{\alpha r}+g^{\alpha s}g^{\beta r})\partial_s\rho_r+(\rho_r-\rho_s)(g^{\beta s}\partial_s g^{\alpha r}+g^{\alpha s}\partial_s g^{\beta r})\\ +Q^\alpha g^{r\beta}+Q^\beta g^{r\alpha}=0 \qquad (\text{sum on } s).\end{aligned} \tag{3.18}$$

Multiplying both sides of (3.18) by $g_{R\alpha}g_{S\beta}$ and summing on α and β we find

$$\begin{aligned}\delta_R^r\partial_S\rho_R+\delta_S^r\partial_R\rho_S+(\rho_r-\rho_S)g_{R\alpha}\partial_S g^{\alpha r}+(\rho_r-\rho_R)g_{S\beta}\partial_R g^{\beta r}\\ +g_{R\alpha}Q^\alpha\delta_S^r+g_{S\beta}Q^\beta\delta_R^r=0.\end{aligned} \tag{3.19}$$

For $R=S=r$ in this expression we find

$$\partial_r\rho_r+g_{r\beta}Q^\beta=0. \tag{3.20}$$

Furthermore, for $r=S$, $r\neq R$ in (3.19) we obtain

$$\partial_R(\rho_S-\rho_R)=(\rho_R-\rho_S)g_{S\beta}\partial_R g^{\beta S}. \tag{3.21}$$

Substitution of (3.20) and (3.21) into (3.18), elimination of all derivative terms $\partial_i\rho_j$ and computation of the coefficient of ρ_s in the resulting equation lead to

$$g_{r\gamma}\partial_s g^{\gamma r}=\partial_s(\ln g^{\alpha r}) \quad \text{if } r\neq s \text{ and } g^{\alpha r}\not\equiv 0. \tag{3.22}$$

Since this expression is independent of α, we can set

$$g^{\alpha r}=B_r^\alpha(x^r)H_r^{-2}. \tag{3.23}$$

Expressions (3.15)–(3.17) and (3.20)–(3.23) lead to

$$\partial_i(\rho_j-\rho_i)=(\rho_j-\rho_i)\partial_i \ln H_j^2, \qquad i,j=1,\cdots,n_1+n_2. \tag{3.24}$$

Comparing this equation with (2.6) we see that the metric $d\hat{s}^2=\sum_{i=1}^{n_1+n_2}H_i^2(dx^i)^2$ is conformal to a Stäckel form metric.

The integrability conditions $\partial_i\partial_j a^{\alpha\beta}=\partial_j\partial_i a^{\alpha\beta}$ for condition 5) are simply that $g^{\alpha\beta}H_{n_1+n_2}^2$ is a Stäckel multiplier for the metric $d\hat{s}^2/H_{n_1+n_2}^2$. Thus, the Hamilton–Jacobi equation separates in the coordinates x. Q.E.D.

Remark 1. It is sufficient to require that condition 5) of Theorem 6 be valid for $i=n_1+1,\cdots,n_1+n_2$ since the requirement that the elements of $\mathcal{K}$ be conformal Killing tensors with $(i,j,k)=(a,\alpha,\beta)$ in (3.14) yields this condition for $i=1,\cdots,n_1$.

Remark 2. Most of the conditions $[A,B]=0$, $A,B\in\mathcal{K}$ (this is just (3.14) with g^{ij} replaced by b^{ij} and $Q^i=0$) are satisfied as a consequence of (3.24) and condition 5). However, the cases $(i,k,l)=(a,a,a)$ and $(i,k,l)=(r,\alpha,\beta)$ lead to the additional requirements

$$\mu_i\partial_i\rho_i=\rho_i\partial_i\mu_i, \qquad i=1,\cdots,m,\ (m+n_1+n_2) \tag{3.25}$$

where A has roots ρ_i and B has roots μ_i.

It is now easy to formulate and prove our main result, the characterization of those involutive families of conformal Killing tensors that correspond to variable separation for the Hamilton–Jacobi equation.

Let $\{x^j\}$ be a local coordinate system on the Riemannian manifold V_n and let $\theta_{(j)}=\lambda_{i(j)}dx^i$, $1\le j\le n$, be a local basis of one-forms on V_n. The dual basis of vector fields is $X^{(h)}=\Lambda^{i(h)}\partial_{x^i}$, $1\le h\le n$, where $\Lambda^{i(h)}\lambda_{i(j)}=\delta^{(h)}_{(j)}$. We say that the forms $\{\theta_{(j)}\}$ are *normalizable* if there exist local analytic functions $g_{(j)}$, y^j such that $\theta_{(j)}=g_{(j)}dy^j$, (no sum).

THEOREM 8. *Suppose there exists a κ-dimensional vector space $\mathcal{Q}$ of second order conformal Killing tensors on V_n such that*:

1) $[A,B]=0$ *for each* $A,B\in\mathcal{Q}$.

2) *There is a basis of one-forms* $\theta_{(h)}=\lambda_{i(h)}dx^i$, $1\le h\le n$, *such that*:

a) *The* n_1 *forms* $\theta_{(a)}$, $1\le a\le n_1$, *are simultaneous eigenforms for every* $A\in\mathcal{Q}$ *with root* ρ_a^A:

$$\left(a^{ij}-\rho_a^A g^{ij}\right)\lambda_{j(a)}=0.$$

b) *The* n_2 *forms* $\theta_{(r)}$, $n_1+1\le r\le n_1+n_2$, *are simultaneous eigenforms for every* $A\in\mathcal{Q}$ *with root* ρ_r^A:

$$\left(a^{ij}-\rho_r^A g^{ij}\right)\lambda_{j(r)}=0.$$

The root ρ_r^A *has multiplicity* 2 *but corresponds to only one eigenform.*

3) $X^{(h)}(\lambda_{i(\alpha)}a^{ij}\lambda_{j(\beta)}-\rho_h^A\lambda_{i(\alpha)}g^{ij}\lambda_{j(\beta)})=0$, $h=n_1+1,\cdots,n_1+n_2$, *for all* $A\in\mathcal{Q}$ *and all* $\alpha,\beta=n_1+n_2+1,\cdots,n$.

4) $[L_\alpha,L_\beta]=0$ *where* $L_\alpha=\Lambda^{i(\alpha)}p_i$ *and each* L_α *is a conformal Killing vector.*

5) $[A,L_\alpha]=0$ *for each* $A\in\mathcal{Q}$.

6) $\kappa=n+n_3(n_3-1)/2$ *where* $n_3=n-n_1-n_2$.

7) $G_{(ab)}\equiv\lambda_{i(a)}g^{ij}\lambda_{j(b)}=0$ *if* $1\le a<b\le n$, *and* $G_{(ar)}=G_{(a\alpha)}=G_{(rs)}=0$ *for* $1\le a\le n_1$, $n_1+1\le r,s\le n_1+n_2$, $n_1+n_2+1\le\alpha\le n$.

Then there exist local coordinates $\{y^j\}$ *for* V_n *such that* $\theta_{(j)}=f^{(j)}(\mathbf{y})dy^j$ *for suitably chosen functions* $f^{(j)}$, *and the Hamilton–Jacobi equation* (1.1) *is separable in these coordinates. Conversely, to every separable coordinate system* $\{y^j\}$ *for the Hamilton–Jacobi equation there corresponds a family* $\mathcal{Q}$ *of conformal Killing tensors on* V_n *with properties* 1)–7).

Proof. This result follows immediately from Theorem 7, once we show that the $\theta_{(h)}$ are normalizable.

The rest of the proof coincides almost word for word with the proof of [8, Thm. 4]. To see this, we remark that the proof of [8, Thm. 4] exploits the relations $[A,B]=0$ for $A,B\in\mathcal{Q}$, identical to those in the present case, and the relations $[A,H]=0$. In the present case, A is only a conformal Killing tensor so $[A,H]=0$ is replaced by (3.14). Multiplying (3.14) by $\lambda_{(m_1)i}\lambda_{(m_2)k}\lambda_{(m_3)l}$ and summing on i,k,l we obtain an identity $E^{A,H}_{m_1,m_2,m_3}$, the right-hand side of which is $\lambda_{(m_1)i}Q^iG_{(m_2m_3)}+\lambda_{(m_2)k}Q^kG_{(m_1m_2)}+\lambda_{(m_3)l}Q^lG_{(m_2m_2)}$. Examining each step in the proof of [8, Thm. 4], we see that the analogy of this identity is needed only in those instances where m_1, m_2, m_3 are such that the right-hand side of $E^{A,H}_{m_1,m_2,m_3}$ vanishes. Q.E.D.

Examples illustrating the practical application of Theorems 4 and 8 can easily be obtained from the corresponding examples in [7] and [8].

REFERENCES

[1] L. P. EISENHART, *Riemannian Geometry*, Princeton Univ. Press, Princeton, NJ, (2nd printing), 1949.

[2] C. P. BOYER, E. G. KALNINS AND W. MILLER, JR., *R-separable coordinates for three-dimensional complex Riemannian spaces*, Trans. Amer. Math. Soc., 242 (1978), pp. 355–376.

[3] E. G. KALNINS AND W. MILLER, JR., *Nonorthogonal R-separable coordinates for four dimensional complex Riemannian spaces*, J. Math. Phys., 22 (1981), pp. 42–50.

[4] P. STÄCKEL, *Über die Integration der Hamilton–Jacobischen Differentialgleichung mittels Separation der Variabelen*, Halle, 1891.

[5] P. MOON AND D. E. SPENCER, *Theorems on separability in Riemannian n-space*, Proc. Amer. Math. Soc., 3 (1952), pp. 635–642.

[6] E. G. KALNINS AND W. MILLER, JR., *R-separation of variables for the four-dimensional flat space Laplace and Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 242 (1978), pp. 355–376.

[7] _____, *Killing tensors and variable separation for Hamilton–Jacobi and Helmholtz equations*, this Journal, 11 (1980), pp. 1011–1026.

[8] _____, *Killing tensors and nonorthogonal variable separation for Hamilton–Jacobi equations*, this Journal, 12 (1981), pp. 617–629.

[9] L. P. EISENHART, *Separable systems of Stäckel*, Ann. of Math. (2), 35 (1934), pp. 284–305.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0010 $01.25/0

ON THE BLOW UP PROBLEM FOR SEMILINEAR HEAT EQUATIONS*

R. O. AYENI†

Abstract. This paper is concerned with the instability behaviour of a differential system arising from the theory of channel and cylindrical flow of viscous fluids with high heat generation. The spatial domain under consideration is the whole space R^n, since usually in this type of problem a bounded domain can be transformed into an unbounded domain. It is shown that if the physical parameter $g(x,t)$, which corresponds to the stress on the fluid is such that $g(x,t)\geq\lambda t^{n+\alpha-1}H(x,t)$, where $H(x,t)$ is the fundamental solution of the heat equation and λ and α are positive constants, then for certain classes of initial conditions the corresponding solution of the initial value problem grows unbounded in a finite time. We obtain an upper bound for the blow up time. We also explain how the method can be applied to a typical physical problem.

1. Introduction.

We consider the problem:

$$\frac{\partial u}{\partial t}-\Delta u=f(x,t,u), \qquad t>0, \quad x\in R^n, \tag{1.1}$$

$$u(x,0)=u_0(x), \qquad u_0(x)\geq 0, \quad x\in R^n, \tag{1.2}$$

$$u(x,t)=0, \tag{1.3}$$

as $|x|\to\infty$.

Problem (1.1)–(1.3) has been investigated by many authors under various conditions on f and u_0. Under certain conditions on f and u_0 (see [3]), there exists a positive local solution $u(x,t)$ of (1.1)–(1.3) such that

(i) $u(x,t)$ is defined in $R^n\times[0,T)$, strictly positive in $R^n\times(0,T)$ and $u(x,0)=u_0(x)$,

(ii) for any $T_0<T$, $u(x,t)$ is bounded and continuous on $R^n\times[0,T_0)$,

(iii) $\partial u/\partial t$ and $\partial^2 u/\partial x_i\partial x_j$ $(1\leq i,j\leq n)$ exist in $R^n\times(0,T)$ and $u(x,t)$ satisfies (1.1)–(1.3) in the classical sense,

where T is a positive number. If T_∞ denotes the supremum of all T satisfying the above three conditions, then the existence of the global solution is the case $T_\infty=\infty$. A positive solution of (1.1)–(1.3) is said to blow up in a finite time and the corresponding T_∞ is called the blow up time of the solution, provided that $T_\infty<\infty$. A global positive solution $u(x,t)$ of (1.1)–(1.3) is said to grow to infinity, if for each positive constant M and each compact set K in R^n there exists $T<\infty$ such that $t>T$ and $x\in K$ imply $u(x,t)>M$.

Kobayahi et al. [3] considered the function f when $f(x,t,\lambda)=f(\lambda)$, $f(\lambda)>0$ for $\lambda>0$, and they gave conditions under which a positive solution of (1.1)–(1.3) blows up in a finite time or grows to infinity. On the other hand, Pao [4, see the literature cited therein] recently studied problem (1.1)–(1.3) when the operator Δ is replaced by a more general elliptic operator L; he not only gave a class of function f for which the solution blows up in a finite time, but also gave an upper bound for the blow up time.

In this paper we investigate another class of functions which arises from the theory of channel and cylindrical flow of viscous fluids with high heat generation (see [6] and [8]). Pao did not cover the class of functions we consider here; moreover, our proofs

*Received by the editors July 10, 1981, and in revised form November 10, 1981.
†Department of Mathematics, University of Ife, Ile-Ife, Nigeria.

depend on the maximum principle while Pao used upper and lower solutions. We consider the function f in the forms

$$f(x,t,u)=g_1(x,t)\exp u, \tag{1.4}$$

$$f(x,t,u)=g_2(x,t)(1+u^2), \tag{1.5}$$

where g_i is a smooth function and

$$g_i(x,t)\geq\delta t^{\alpha-1}\exp\left(-\frac{|x|^2}{4t}\right), \qquad i=1,2, \tag{1.6}$$

$$g_i(\pm\infty,t)=0, \qquad i=1,2, \tag{1.7}$$

and α and δ are positive constants.

2. The main result.

LEMMA. *Let $f\equiv g_i$ $(i=1,2)$ such that g_i satisfies* (1.6) *and* (1.7). *Then the unique solution u of* (1.1)–(1.3) *satisfies*

$$u(x,t)\geq\delta\left(\frac{\alpha+n}{2}\right)^{-1}t^{\alpha}\exp\left(-\frac{|x|^2}{4t}\right).$$

THEOREM 1. *Let f satisfy* (1.4), (1.6) *and* (1.7). *The unique solution u of* (1.1)–(1.3) *satisfies*

(i) $$u(x,t)\geq\log\left[1-\delta\left(\alpha+\frac{n}{2}\right)^{-1}t^{\alpha}\exp\left(-\frac{|x|^2}{4t}\right)\right]^{-1},$$

(ii) $$u(0,t)\to\infty \quad \text{as } t\to\left[\delta^{-1}\left(\frac{\alpha+n}{2}\right)\right]^{1/\alpha}$$

THEOREM 2. *Let f satisfy* (1.5), (1.6) *and* (1.7). *Then the unique solution u of* (1.1)–(1.3) *satisfies*

(i) $$u(x,t)\geq\tan\delta\left(\alpha+\frac{n}{2}\right)^{-1}t^{\alpha}\exp\left(-\frac{|x|^2}{4t}\right),$$

(ii) $$u(0,t)\to\infty \quad \text{as } t\to\left[\pi\delta^{-1}\frac{(\alpha+n/2)}{2}\right]^{1/\alpha}.$$

3. Proofs of the lemma and the theorems. We will not bother to prove the existence and uniqueness of the solution since the method of the proof is well documented in the literature (see for example [4]). Furthermore, we will take $u_0(x)\equiv 0$. The results are true if $u_0(x)>0$.

Proof of the lemma. Let $w(x,t)=u(x,t)-\delta(\alpha+n/2)^{-1}t^{\alpha}\exp(-|x|^2/4t)$. Then

$$\frac{\partial w}{\partial t}-\Delta w\geq 0, \tag{3.1}$$

$$w(x,0)=0, \tag{3.2}$$

$$w(\pm\infty,t)=0. \tag{3.3}$$

By the maximum principle for parabolic equations (see [7, p. 183]) the solution $w(x,t)$ of (3.1)–(3.3) satisfies $w(x,t)\geq 0$. The result follows.

Proof of Theorem 1. Let $v=1-e^{-u}$. Then $u=\log(1-v)^{-1}$. This is increasing in v if $0<v<1$ and v satisfies

$$\frac{\partial v}{\partial t}-\Delta v\geq g_1(x,t), \tag{3.4}$$

$$v(x,0)=0, \tag{3.5}$$

$$v(\pm\infty,t)=0. \tag{3.6}$$

Hence, by the lemma, $v(x,t)\geq\delta(\alpha+n/2)^{-1}t^{\alpha}\exp(-|x|^2/4t)$. Then

(i) $$u(x,t)=\log(1-v(x,t))^{-1},$$

(ii) $$u(0,t)\to\infty \quad \text{as } t\to\left[\delta^{-1}\left(\alpha+\frac{n}{2}\right)\right]^{1/\alpha}.$$

Proof of Theorem 2. Let $v=\tan^{-1}u$. Then $u=\tan v$. This is increasing in v if $0<v<\pi/2$ and v satisfies

$$\frac{\partial v}{\partial t}-\Delta v\geq g_2(x,t), \tag{3.7}$$

$$v(x,0)=0, \tag{3.8}$$

$$v(\pm\infty,t)=0. \tag{3.9}$$

Hence by the lemma, $v(x,t)\geq\delta(\alpha+n/2)^{-1}t^{\alpha}\exp(-|x|^2/4t)$. Then

(i) $$u(x,t)=\tan v(x,t),$$

(ii) $$u(0,t)\to\infty \quad \text{as } t\to\left[\pi\delta^{-1}\frac{\alpha+n/2}{2}\right]^{1/\alpha}.$$

4. Applications. When $g_i(x,t)=\delta/(x+1)^4$, $x>0$, let

$$w=v-\left[\frac{\delta t}{24(1+x)^4}\right]\left[\exp\left(-\frac{(x-1)^2}{4t}\right)-\exp\left(-\frac{(x+1)^2}{4t}\right)\right].$$

Then

$$\frac{\partial w}{\partial t}-\frac{1}{(1+x)}\frac{\partial w}{\partial x}-\frac{\partial w}{\partial x^2}\geq 0, \tag{4.1}$$

$$w(x,0)=0, \tag{4.2}$$

$$w(0,t)=w(\infty,t)=0. \tag{4.3}$$

By the maximum principle, $w(x,t)\geq 0$.

Remark. In the case of a bounded domain $\Omega\subset R^n$, the method of Pao [4] is applicable if $g_i(x,t)\geq\delta$.

Acknowledgment. This paper has benefited from correspondence with Dr. H. R. Dowson. The author also wishes to thank the referee for thoughtful comments which improved the original version of this paper.

REFERENCES

[1] R. O. AYENI, Ph. D. Thesis, Cornell Univ., Ithaca, NY, 1978.

[2] R. O. AYENI AND G. S. S. LUDFORD, *Nonexistence of global solution of a boundary value problem of parabolic type*, Nigerian J. Natural Sciences C, to appear.

[3] K. KOBAYASHI, T. SIRAO AND H. TANAKA, *On the growing up problem for semilinear heat equations*, J. Math. Soc. Japan, 29 (1977), pp. 407–424.

[4] C. V. PAO, *Nonexistence of global solutions for an integrodifferential system in reactor dynamics*, This Journal, 11 (1980), pp. 559–564.

[5] 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.

[6] J. R. A. PEARSON, *Variable-viscosity flows in channels with high heat generation*, J. Fluid Mech., 83 (1977), pp. 191–206.

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

[8] A. M. STOLIN, S. A. BOSTANDZHIYAN AND N. Y. PLOTNIKOV, *Conditions for occurrence of hydrodynamic thermal explosion in flow of power-law fluids*, Heat transfer-Soviet-Research, 10, 1(1978), pp. 86–93.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0011 $ 01.25/0

QUALITATIVE BEHAVIOR AND BOUNDS IN A NONLINEAR PLASMA PROBLEM*

CATHERINE BANDLE† AND RENÉ P. SPERB‡

Abstract. Bounds for the boundary values of the variational solutions of a plasma problem are constructed, estimates for the distance between the boundary of the domain and the region filled with plasma are given and an isoperimetric inequality for the area of this region is derived.

1. Introduction.

We shall consider problems of the form

$$\begin{aligned} &\Delta u+\lambda(u^+)^p=0 \quad \text{in } D\subset\mathbb{R}^2, \qquad u^+:=\max(u,0),\\ &u=\alpha \quad \text{on } \partial D,\\ &-\oint_{\partial D}\frac{\partial u}{\partial n}ds=I, \end{aligned} \tag{1.1}$$

where $p>1$ and $\frac{\partial}{\partial n}$ is the outer normal derivative of D; $\lambda>0$ and $I>0$ are given and α is an unknown constant determined by the data. This equation has its origin in plasma physics and describes the equilibrium of a plasma confined in a Tokomak machine. The plasma occupies the unknown region $D^+:=\{x: u(x)>0\}$ and $D^-:=\{x: u(x)<0\}$ is the vacuum. The model case $p=1$ has been treated by Temam [14], [15] and problem (1.1) as well as some of its generalizations have been investigated by Berestycki and Brézis [4]. They proved the existence of a solution, called a *variational solution*, which also solves the problem

$$\begin{aligned} &J[v]:=\int_D\left[|\nabla v|^2-\frac{2\lambda}{p+1}\{v^+\}^{p+1}\right]dx+2Iv(\partial D)\to\text{infimum},\\ &v\in K:=\left\{w\in H_0^1(D)\oplus\mathbb{R}:\lambda\int_D(w^+)^p dx=I\right\}. \end{aligned} \tag{1.2}$$

Very little is known on the number of variational solutions. A result of Berestycki and Brézis [4, Thm. 4] and estimates in [3] suggest that for small α and I there is a unique solution. For the model case $p=1$ examples are known with several solutions [5], [11].

It follows immediately from the maximum principle that $u(x)>\alpha$ in D. If $\alpha<0$, then D^+ is completely contained in D and the problem (1.1) can be interpreted as a free boundary value problem with ∂D^+ as *free boundary*. Hence a free boundary appears if and only if α is negative. It was shown in [3] that in the case of variational solutions the sign of α depends only on I. More precisely, the following statement holds: There exists a number $I_s>0$ such that

$$\begin{aligned} &\alpha>0 \quad \text{for all } I<I_s,\\ &\alpha<0 \quad \text{for all } I>I_s. \end{aligned}$$

For $I=I_s$ we have either $\alpha=0$ if there is only one variational solution, or if there are at least two variational solutions, one is with $\alpha\geq 0$ and one with $\alpha\leq 0$.

* Received by the editors March 17, 1981, and in revised form February 15, 1982.
† Mathematisches Institut der Universität Basel, CH-4051 Basel, Switzerland.
‡ Seminar für Angewandte Mathematik, Eidgen Tech Hochschule, CH-8092 Zürich, Switzerland.

Let us now consider problem (1.1) in the circle $D^*=\{x:|x|<R\}$. According to a result of Gidas, Ni and Nirenberg [6] the solutions are radially symmetric. Denote by u_0 the unique *positive* solution of

$$\Delta u_0+\lambda u_0^p=0 \quad \text{in } D^*, \qquad u_0=0 \quad \text{on } \partial D^*. \tag{1.3}$$

For $p>1$ the existence is guaranteed by a result of Levinson [8], and the uniqueness follows from an argument in [6]. We have

LEMMA 1.1. (i) *The circle has a unique solution.*

(ii) *A free boundary appears if and only if*

$$I_0 := \oint_{\partial D}\left|\frac{\partial u_0}{\partial n}\right| ds<I.$$

Proof. Any radially symmetric solutions of $\Delta u+\lambda u^p=0$, which is regular at the origin, can be expressed in terms of u_0. Thus for $\alpha>0$ we have

$$u(r)=C^{2/(p-1)}u_0(Cr) \qquad (r:=|x|),$$

C being a constant smaller than one such that

$$I=\oint_{\partial D^*}\left|\frac{\partial u}{\partial n}\right| ds=-2\pi RC^{[2/(p-1)]+1}u_0'(CR)$$

$$=C^{2/(p-1)}\oint_{\{|x|=CR\}}\left|\frac{\partial u_0}{\partial n}\right| ds=C^{2/(p-1)}\lambda\int_{\{|x|<CR\}}u_0^p\,dx.$$

The right-hand side depends monotonically on C. There exists therefore a unique solution provided that $I<I_0$. If $\alpha<0$ we have

$$u(r)=\begin{cases} C^{2/(p-1)}u_0(Cr) & \text{in } D^{*+}:=\{x:|x|<\rho\},\\ A\log r+B & \text{in } D^{*-}:=\{x:\rho<|x|<R\},\end{cases}$$

where $u(\rho)=0$ implies that $B=-A\log\rho$ and A is determined by the continuity assumption for $u'(r)$ at $r=\rho$, that is,

$$C^{[2/(p-1)]+1}u_0'(C\rho)=\frac{A}{\rho}.$$

Moreover we must have

$$I=\oint_{\partial D^{*+}}\left|\frac{\partial u}{\partial n}\right| ds=-2\pi\rho C^{[2/(p-1)]+1}u_0'(C\rho)=C^{2/(p-1)}\oint_{\partial D^*}\left|\frac{\partial u_0}{\partial n}\right| ds, \tag{1.4}$$

which determines a unique $C>1$. The lemma is now obvious. □

The aim of this paper is to obtain various estimates for the solutions of (1.1), depending only on the area A of D. By means of an isoperimetric inequality of Payne, Sperb and Stakgold [10] and Schwarz's symmetrization [1] we show:

Among all domains of given area the boundary value of a variational solution is smallest for the circle.

We then study the geometry of D^+ and prove for variational solutions:

Among all domains of given total area, the area of D^+ achieves its minimum for the circle.

The last part contains mainly applications of a gradient bound from Payne and Stakgold [9]; cf. also [12]. First we give an estimate for the location of D^+ and then make use of level line techniques [1], [12] to get relations between $u_{\max}$ and A.

For the case $p=1$ similar investigations were carried out in [2]. Other methods were used in [5] to obtain information on the asymptotic behavior of the variational solutions.

2. Inequalities.

2.1. The results of this section are based on an inequality by Payne, Sperb and Stakgold [10], which for the solutions u of problem (1.1) may be written as

$$\lambda\int_{D^+}u^{p+1}\,dx\leq\begin{cases}\lambda A\alpha^{p+1}+\dfrac{p+1}{8\pi}I^2 & \text{if } \alpha>0,\\[2ex] \dfrac{p+1}{8\pi}I^2 & \text{if } \alpha\leq 0.\end{cases}\tag{2.1}$$

Equality holds in both cases for the circle.

2.2. Upper bounds for $u_{\max}$. This inequality gives rise to an estimate for $u_{\max}$ in the case where $\alpha<0$. Denoting by $g(x,y)$ the Green's function of the Laplacian with $g(x,\cdot)=0$ on ∂D, we have for any solution of (1.1)

$$u(x)=\lambda\int_D g(x,y)[u^+(y)]^p dy+\alpha.$$

By means of Hölder's inequality it follows that

$$|u(x)-\alpha|\leq\lambda\left\{\int_D g^{p+1}(x,y)\,dy\right\}^{1/(p+1)}\left\{\int_{D^+}u^{p+1}\,dy\right\}^{p/(p+1)}$$

In view of (2.1) we conclude

LEMMA 2.1. *If u is any solution of* (1.1) *with $\alpha\leq 0$, then*

$$u_{max}\leq\left(\frac{p+1}{8\pi}I^2\right)^{p(p+1)}\max_{x\in\bar{D}}\left\{\lambda\int_D g^{p+1}(x,y)\,dy\right\}^{1/(p+1)}$$

Remarks. (1) Estimates for $\max_{x\in 0}\{\int_D g^{p+1}(x,y)\,dy\}^{1/(p+1)}$ are given in [14], [1]. There, an isoperimetric upper bound is given, which depends only on the area A of D and p, and which tends to zero as $A\to 0$.

(2) For $p=1$ it has been shown in [2] that among all domains of given area, $u_{\max}$ of a variational solution achieves its maximal value for the circle. This statement can be expressed as

$$u_{\max}\leq\begin{cases}\dfrac{I}{2\pi\sqrt{\lambda}\,RJ_1\left(\sqrt{\lambda}\,R\right)} & \text{if } R\leq j_0/\sqrt{\lambda},\\[2ex] \dfrac{I}{2\pi j_0 J_1(j_0)} & \text{otherwise,}\end{cases}$$

where R is the radius of the circle with the same area as D and J_n is the Bessel function of order n and $j_0=2.4048\cdots$ is the first zero of J_0.

2.3. Upper bounds for α. We first note that for any solution of (1.1)

$$J[u]=I\alpha+\frac{p-1}{p+1}\lambda\int_{D^+}u^{p+1}\,dx.\tag{2.2}$$

This relation holds independently of the sign of α. If we restrict ourselves to variational solutions, it follows that for any $v \in K$

$$J[v] \geq I\alpha + \frac{p-1}{p+1}\lambda \int_{D^+} u^{p+1}\,dx. \tag{2.3}$$

From Hölder's inequality and the flux condition we conclude that

$$\lambda \int_{D^+} u^{p+1}\,dx \geq \left(\frac{I^{p+1}}{\lambda A^+}\right)^{1/p} \geq \left(\frac{I^{p+1}}{\lambda A}\right)^{1/p}, \qquad A^+ := \int_{D^+} dx. \tag{2.4}$$

Inequalities (2.3) and (2.4) yield

LEMMA 2.2. *If $v \in K$ is any admissible function in the variational characterization* (1.2), *then the boundary values α of a variational solution satisfy*

$$\alpha \leq I^{-1}\left[J[v] - \frac{p-1}{p+1}\left(\frac{I^{p+1}}{\lambda A}\right)^{1/p}\right],$$

A being the area of D.

2.4. A lower bound for α. Let u be any solution of (1.1) in D and let u^* be the solution in the circle of the same area as D, say D^*. By means of the Schwarz symmetrization it follows that

$$J_D[u] \geq J_{D^*}[u^*]. \tag{2.5}$$

From (2.2) and inequality (2.1) we find

$$J_{D^*}[u^*] \leq I\alpha + \frac{p-1}{p+1}\begin{cases} \lambda A\alpha^{p+1} + \dfrac{p+1}{8\pi}I^2 & \text{if } \alpha > 0, \\ \dfrac{p+1}{8\pi}I^2 & \text{otherwise.} \end{cases} \tag{2.6}$$

Moreover, we have

$$J_{D^*}[u^*] = I\alpha^* + \frac{p-1}{p+1}\begin{cases} \lambda A(\alpha^*)^{p+1} + \dfrac{p+1}{8\pi}I^2 & \text{if } \alpha^* > 0, \\ \dfrac{p+1}{8\pi}I^2 & \text{otherwise.} \end{cases} \tag{2.7}$$

From these observations we conclude

THEOREM 2.1. *For any solution u of problem* (1.1) *we have*

$$\alpha \geq \alpha^*,$$

where α^* is the boundary value of the unique solution in D^*.

Proof. The statement of the theorem is a consequence of (2.5), (2.6) and (2.7). We have to distinguish between the cases.

Case 1. $\alpha^* \leq 0$, $\alpha \leq 0$. Then

$$I\alpha^* + \frac{p-1}{8\pi}I^2 \leq I\alpha + \frac{p-1}{8\pi}I^2,$$

which proves that $\alpha^* \leq \alpha$.

Case 2. $\alpha^* \leq 0$, $\alpha > 0$. The inequality of Theorem 2.1 is then trivially implied.

Case 3. $\alpha^* \geq 0$, $\alpha \geq 0$. Then (2.5), (2.6) and (2.7) yield

$$I\alpha^* + \frac{p-1}{p+1}\lambda A(\alpha^*)^{p+1} \leq I\alpha + \frac{p-1}{p+1}\lambda A\alpha^{p+1}.$$

Since we assumed that $p>1$, the function $f(t):=It+((p-1)/(p+1))\lambda A t^{p+1}$ is monotone, and we therefore have $\alpha^*\leq\alpha$.

Case 4. $\alpha^*\geq 0$, $\alpha<0$. By (2.5), (2.6) and (2.7) we have

$$I\alpha^*+\frac{p-1}{p+1}\lambda A(\alpha^*)^{p+1}\leq I\alpha<0.$$

contradicting our assumption $\alpha^*>0$. The proof of the theorem is thus completed.

Remarks. For $p=1$ this result was already found in [2]. If $p=1$, all variational solutions have the same value of α, and all other solutions have larger boundary values [2]. It is not yet clear whether a similar relation holds for general p.

2.5. Remarks on the variational solutions. Let u be a variational solution with $\alpha<0$. In this case there is a free boundary and we have $D^+\subset D$. As Berestycki and Brézis observed in [4], D^+ is simply connected. This is a direct consequence of the minimum property of the variational solutions.

Let us now consider the following variational problem:

$$F[v]=\frac{\int_{D^+}|\nabla v|^2dx}{\left(\int_{D^+}v^{p+1}dx\right)^{2/(p+1)}}\to\text{infimum}\quad\text{over}$$

$$v\in Q(D^+):=\left\{w\in H_0^1(D^+):\lambda\int_{D^+}w^{p+1}dx=\int_{D^+}|\nabla w|^2dx\right\}. \tag{2.8}$$

By standard methods using a Sobolev inequality and compactness arguments, a solution of (2.8) is easily established. Notice that Levinson [8] considered similar problems in order to prove the existence of positive solutions for nonlinear Dirichlet problems. He showed that if ∂D^+ is sufficiently smooth, say of class C^1, then the solutions of (2.8) solve also the problem

$$\Delta u+\lambda u^p=0\quad\text{in } D^+,\qquad u=0\quad\text{on }\partial D^+.$$

As a first step we show

LEMMA 2.3. *A variational solution u restricted to D^+ is a solution of* (2.8).

Proof. Suppose that u is not a solution of (2.8). Then

$$\int_{D^+}|\nabla u|^2dx>\int_{D^+}|\nabla v_0|^2dx,$$

for any solution v_0 of (2.8).

Let

$$w=\begin{cases}\beta v_0 & \text{in } D^+,\\ u & \text{in } D-D^+,\end{cases}$$

where $\beta>0$ is determined such that $w\in K$, that is, w is admissible for problem (1.2). Then

$$J[w]=I\alpha+\beta^2\int_{D^+}|\nabla v_0|^2dx-\frac{2\beta^{p+1}}{p+1}\lambda\int_{D^+}v_0^{p+1}dx$$

$$=I\alpha+\int_{D^+}|\nabla v_0|^2dx\left\{\beta^2-\frac{2\beta^{p+1}}{p+1}\right\}.$$

Since for $\beta>0$ we have

$$\beta^2-\frac{2\beta^{p+1}}{p+1}\leq\frac{p-1}{p+1},$$

it follows that

$$J[w]\leq I\alpha+\frac{p-1}{p+1}\int_{D^+}|\nabla v_0|^2dx<I\alpha+\frac{p-1}{p+1}\int_{D^+}|\nabla u|^2dx=J[u],$$

which contradicts the minimum property of u. The assertion of the lemma thus follows. □

Let D' be the circle of the same area as D^+ and let, for any domain B, $\mu(B):=\inf_{Q(B)}F[v]$. This quantity is uniquely determined although the solutions of (2.8) may not be unique.

LEMMA 2.4. *Under the assumptions stated above we have*

$$\mu(D')\leq\mu(D^+).$$

Proof. We shall prove this lemma by means of Schwarz symmetrization. Let u_s: $D'\to R^+$ be the radially symmetric function such that

$$\text{area}\{x\in D^+: u(x)>t\}=\text{area}\{x\in D': u_s(x)>t\}\quad\forall t\in(0,u_{\max}].$$

Since $u(x)$ is real analytic in D^+, u_s belongs to $H_0^1(D')$ and satisfies [1]

$$\int_{D'}|\nabla u_s|^2dx\leq\int_{D^+}|\nabla u|^2dx$$

and

$$\int_{D'}u_s^{p+1}dx=\int_{D^+}u^{p+1}dx.$$

Let β be such that $\beta u_s\in Q(D')$. Since $\beta\leq 1$, it follows that

$$\mu(D')\leq\left(\int_{D'}|\nabla\beta u_s|^2dx\right)^{(p-1)/(p+1)}\leq\left(\int_{D^+}|\nabla u|^2dx\right)^{(p-1)/(p+1)}=\mu(D^+),$$

which establishes the lemma. □

If u' denotes any solution of (2.8) in D', we conclude from the previous lemma and (2.1) that

$$\lambda\int_{D'}(u')^{p+1}dx\leq\lambda\int_{D^+}u^{p+1}dx\leq\frac{p+1}{8\pi}I^2. \tag{2.9}$$

As already mentioned, u' solves $\Delta u'+\lambda(u')^p=0$ in D', $u'=0$ on $\partial D'$. Hence by (2.1)

$$\lambda\int_{D'}(u')^{p+1}dx=\frac{p+1}{8\pi}\left\{\lambda\int_{D'}(u')^p dx\right\}^2. \tag{2.10}$$

Combining this identity with (2.9) we find that

$$I':=\lambda\int_{D'}(u')^p dx\leq I. \tag{2.11}$$

Let u_0 and I_0 be defined as in Lemma 1.1. We are now in a position to prove:

THEOREM 2.2. *Let D^* be the circle with the same area as D and suppose that $\alpha<0$. Denote by u a variational solution of* (1.1) *in D and let u^* be the solution in D^*. Furthermore let*

$$D^+=\{x\in D: u(x)>0\}\quad\text{and}\quad D^{*+}=\{x\in D^*: u^*(x)>0\}.$$

Then the following relation holds for the areas of D^+ and D^{*+}:

$$A(D^+)\geq A(D^{*+}).$$

Proof. Let ρ and ρ' be the radii of D^{*+} and D', respectively. In view of (1.4) it follows that $\rho = R(I_0/I)^{(p-1)/2}$, and hence ρ decreases if I increases. By the same argument we have $\rho' = R(I_0/I')^{(p-1)/2}$, and in view of (2.11) we get $\rho' > \rho$, which implies that $A(D^+) = \pi(\rho')^2 \geq \pi\rho^2 = A(D^{*+})$. □

Remarks. (1) If $I \to \infty$, then $A(D^{*+}) \to 0$.

(2) $A(D^{*+})$ does not depend on the size of D^*, but only on I.

3. Applications of a theorem by Payne and Stakgold.

3.1. Some direct consequences. Throughout this section we shall assume D to be *convex.*

The following result can be found in [9].

THEOREM 3.1. (Payne and Stakgold). *Let u be a solution of* (1.1) *and D a convex domain. Then one has*

$$|\nabla u|^2 + \frac{2\lambda}{p+1}\{u^+\}^{p+1} \leq \frac{2\lambda}{p+1} u_{\max}^{p+1} \quad \text{in } D. \tag{3.1}$$

Inequality (3.1) will be the basis for the derivation of a number of inequalities in problem (1.1). The techniques used below have been described in Payne, Sperb and Stakgold [10] and need only some modification in the present context. Our general aim is to derive some inequalities relating the geometry of D and $u_{\max}$, α and the total flux. It should be pointed out that our considerations hold for any solution, and are not restricted to the variational ones.

From now on we consider only problems with a free boundary, that is, $\alpha < 0$.

Let x_M be a point of D^+ where u assumes the value $u_{\max}$ and y be any point of ∂D. Denote by g the straight line joining x_M to y, and let r measure the distance from a point on g to x_M. Putting $\frac{d}{dr}$ for the derivative along g, we get from inequality (3.1)

$$-\frac{du}{dr} \leq \left\{ \frac{2\lambda}{p+1}\left(u_{\max}^{p+1} - u^{p+1}\right) \right\}^{1/2} \quad \text{in } D^+, \tag{3.2}$$

$$-\frac{du}{dr} \leq \left\{ \frac{2\lambda}{p+1} u_{\max}^{p+1} \right\}^{1/2} \quad \text{in } D^-. \tag{3.3}$$

For the distance from x_M to the first intersection point z on ∂D^+, (3.2) yields

$$\int_0^{u_{max}} \frac{du}{\left(u_{\max}^{p+1} - u^{p+1}\right)^{1/2}} \leq \left(\frac{2\lambda}{p+1}\right)^{1/2} |z - x_M|. \tag{3.4}$$

Now the integral in (3.4) can be calculated explicitly giving

$$u_{\max}^{(1-p)/2} N(p) \leq \left(\frac{2\lambda}{p+1}\right)^{1/2} |z - x_M| \tag{3.5}$$

with

$$N(p) := \sqrt{\pi}\, \frac{\Gamma\left(1 + \frac{1}{p+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{p+1}\right)}.$$

Since this inequality is independent of the choice of g, it holds in particular for the point z nearest to x_M. (3.5) then yields

LEMMA 3.1. *Let*

$$d := u_{\max}^{(1-p)/2}\sqrt{\frac{p+1}{2}}\,N(p).$$

Then

$$A^+ \geq \pi d^2 \qquad (A^+ := A(D^+)). \tag{3.6}$$

The upper bound for $u_{\max}$ of §2.2. can be used to get an estimate from below for d.

Now let $y \in \partial D$ and $z \in \partial D^+$ be two points such that $\operatorname{dist}(\partial D, \partial D^+) = |y-z|$ and let g be the straight line joining y to z. It is clear that (3.3) holds also in this case if r denotes the distance from z. An integration yields

LEMMA 3.2. *The distance δ between ∂D^+ and ∂D is bounded from below by*

$$\delta \geq \left(\frac{p+1}{2\lambda}\right)^{1/2} |\alpha| u_{\max}^{-(p+1)/2}. \tag{3.7}$$

We can again combine this result with Lemmas 2.1 and 2.2 to construct an estimate which depends only on certain geometrical data of D.

Remark. Related results for $p=1$ are found in [5]. Caffarelli and Friedman used different methods which enabled them to study the asymptotic behavior of D^+ as $\lambda \to \infty$.

3.2. Combination of (3.1) and integration along level lines. It is possible to derive yet another set of inequalities in problem (1.1) if the inequality (3.1) is used in conjunction with techniques described in Payne, Sperb and Stakgold [10].

Let $D(t)$ be the subdomain of D where $u > t$ and let $\Gamma(t)$ be its boundary. Denote by $a(t)$ the area enclosed by $\Gamma(t)$, and set

$$E(t) := \int_{D(t)} u^p \, dx = \frac{1}{\lambda} \oint_{\Gamma(t)} |\nabla u| \, ds. \tag{3.8}$$

$\Gamma(t)$ is real analytic for $t \neq 0$ and it follows therefore that $a(t)$ is monotonic. Let $t(a)$ be its inverse. In addition, it is known [1] that $t(a)$ is differentiable almost everywhere. Routine level technique yields

$$-\left(\frac{dt}{da}\right)^{-1} = \oint_{\Gamma(t)} \frac{1}{|\nabla u|}\, ds,$$

whence

$$-\frac{dt}{da}\lambda E(a) = \left(\oint_{\Gamma(t)} \frac{ds}{|\nabla u|}\right)^{-1} \oint_{\Gamma(t)} |\nabla u| \, ds. \tag{3.9}$$

Inequality (3.9) can now be combined with (3.1) to give the two inequalities

$$-\frac{dt}{da}E(a) \leq \frac{2}{p+1}\left(u_{\max}^{p+1} - (u(a))^{p+1}\right), \qquad a \in (0, A^+), \tag{3.10}$$

$$-\frac{dt}{da}E(a) = -\frac{du}{da} I \leq \frac{2}{p+1} u_{\max}^{p+1}, \qquad a \in (A^+, A). \tag{3.11}$$

We first integrate inequality (3.10) in two steps. We multiply both sides in (3.10) by $(t(a))^p = dE/da$ and separate variables to obtain

$$\frac{-(p+1)t^p(dt/da)}{u_{\max}^{p+1}-t^{p+1}} \leq \frac{2(dE/da)}{E}.$$

Integration from a to A^+ yields

$$\frac{u_{\max}^{p+1}}{u_{\max}^{p+1}-t^{p+1}} \leq \left(\frac{E(0)}{E}\right)^2 = \left(\frac{I}{\lambda E}\right)^2.$$

Hence

$$u_{\max}^p\left(1-\left(\frac{\lambda E}{I}\right)^2\right)^{p/(p+1)} \geq t^p = \frac{dE}{da},$$

which finally leads to

$$u_{\max}^p A^+ \geq \int_0^{I/\lambda} \frac{dE}{\left(1-\left(\frac{\lambda E}{I}\right)^2\right)^{p/(p+1)}} = \frac{I}{\lambda}\int_0^1 \frac{dy}{(1-y^2)^{p/(p+1)}}.$$

We state the result as

LEMMA 3.3. *If*

$$M(p) := 2^{-2p/(p+1)} \frac{\left[\Gamma\left(\frac{1}{p+1}\right)\right]^2}{\Gamma\left(\frac{2}{p+1}\right)},$$

then

$$u_{\max}^p A^+ \geq \frac{I}{\lambda} M(p).$$

Complements. (1) Integration of (3.11) yields

$$|\alpha| I > \left[\frac{2}{p+1}\right] u_{\max}^{p+1}(A-A^+).$$

(2) Another interesting combination of inequalities is possible as follows. For $a \in (A^+, A)$ we may write (3.9) as

$$-\frac{da}{du}\lambda E = -\frac{da}{du} I = \oint_{\Gamma(t)} \frac{ds}{|\nabla u|} \oint_{\Gamma(t)} |\nabla u|\, ds \geq 4\pi a,$$

where Schwarz's inequality and the classical isoperimetric inequality were used in the last step. Thus we have

$$\frac{du}{da} + \frac{I}{4\pi a} \geq 0, \qquad a \in (A^+, A),$$

and after integration

$$|\alpha| < \frac{I}{4\pi} \log \frac{A}{A^+},$$

which is essentially Carleman's inequality [1].

Acknowledgment. We would like to thank the referee for pointing out several errors and for having drawn our attention to the paper of Caffarelli and Friedman.

REFERENCES

[1] C. BANDLE, *Isoperimetric Inequalities and Applications*, Pitman Publ., London, 1980.
[2] _____, *Abschätzungen der Randwerte bei nichtlinearen elliptischen Gleichungen aus der Plasmaphysik*, ISNM (Birkhäuser), 56 (1981), pp. 1–17.
[3] C. BANDLE AND M. MARCUS, *On the boundary values of solutions of a problem arising in plasma physics*, Technion Preprint Ser. MT-528, 1982.
[4] H. BERESTYCKI AND H. BRÉZIS, *On a free boundary value problem arising in plasma physics*, Nonlinear Anal., 4 (1980), pp. 415–436.
[5] L. A. CAFFARELLI AND A. FRIEDMAN, *Asymptotic estimates for the plasma problem*, Duke Math. J., 47 (1980), pp. 705–742.
[6] B. GIDAS, W. NI AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.
[7] G. KEADY AND J. NORBURY, *A semilinear elliptic eigenvalue problem*, II. *The Plasma Problem*, Proc. Roy. Soc. Edinburgh, A, 87 (1980), pp. 83–109.
[8] N. LEVINSON, *Positive eigenfunctions for* $\Delta u+\lambda f(u)=0$, Arch. Rational Mech. Anal., 11 (1962), pp. 258–272.
[9] L. E. PAYNE AND I. STAKGOLD, *Nonlinear problems in nuclear reactor analysis*, Lecture Notes in Mathematics 322, Springer, New York, 1972, pp. 298–307.
[10] L. E. PAYNE, R. P. SPERB AND I. STAKGOLD, *On Hopf type maximum principles for convex domains*, Nonlinear Anal., 1 (1977), pp. 547–559.
[11] D. G. SCHAEFFER, *Non-uniqueness in the equilibrium shape of a confined plasma*, Comm. Partial Differential Equations, 2 (1977) , pp. 587–600.
[12] R. P. SPERB, *Maximum Principles and their Applications*, Academic Press, New York, 1981.
[13] I. STAKGOLD, *Gradient bounds for plasma confinement*, Math. Meth. Appl. Sci., 2 (1980), pp. 68–72.
[14] R. TEMAM, *A nonlinear eigenvalue problem: the shape at equilibrium of a confined plasma*, Arch. Rational Mech. Anal., 60 (1975), pp. 51–73.
[15] _____, *Remarks on a free boundary value problem arising in plasma physics*, Comm. Partial Differential Equations, 2 (1977), pp. 563–585.
[16] H. F. WEINBERGER, *Symmetrization in uniformly elliptic problems*, Studies in Mathematical Analysis and Related Topics, Stanford Univ., Stanford, CA., 1962, pp. 424–428.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0012 $01.25/0

EIGENVALUES OF POSITIVE DEFINITE KERNELS*

J. B. READE†

Abstract. The main result is that the eigenvalues of any continuously differentiable positive definite kernel are $o(1/n^2)$. The method of proof is to approximate the kernel by kernels of finite rank in such a way that the difference is positive definite. The tail of the eigenvalue series can then be related to the trace integral of the difference by means of Mercer's theorem and a trace norm version of the Weyl–Courant minimax principle. It is conjectured that, for p times continuously differentiable positive definite kernels, the eigenvalues are $o(1/n^{p+1})$.

1. Introduction. Suppose that $K(x,t)\in L^2[0,1]^2$ and

$$K(t,x)=\overline{K(x,t)}$$

for almost all $0\le x,t\le 1$. Then the operator T defined by

$$Tf(x)=\int_0^1 K(x,t)f(t)\,dt$$

is a compact symmetric operator on $L^2[0,1]$. It therefore has an infinite sequence (λ_n) of real eigenvalues which converges to zero. (See e.g. [1, p. 233].) H. Weyl has proved (see [2]) that, if these eigenvalues are arranged in decreasing order of modulus, then for $K(x,t)\in C^1[0,1]^2$ we have

$$\lambda_n=o\left(\frac{1}{n^{3/2}}\right)$$

as $n\to\infty$. It is the purpose of this paper to show that, if $K(x,t)$ is also assumed to be positive definite, i.e.

$$\int_0^1\int_0^1 K(x,t)f(x)\overline{f(t)}\,dx\,dt\ge 0$$

for all $f\in L^2[0,1]$, then Weyl's result can be improved to

$$\lambda_n=o\left(\frac{1}{n^2}\right)$$

as $n\to\infty$.

2. Best approximations. If (ϕ_n) are the corresponding eigenfunctions of T, then they can be assumed to form an orthonormal sequence. (See [1, p. 234].) Also the operator T and its kernel $K(x,t)$ have the following eigenfunction expansions:

$$Tf=\sum_1^\infty \lambda_n\langle f,\phi_n\rangle\phi_n,\qquad K(x,t)=\sum_1^\infty \lambda_n\phi_n(x)\overline{\phi_n(t)},$$

convergent in mean square over $[0,1]$ and $[0,1]^2$ respectively, where $\langle f,\phi_n\rangle$ denotes the inner product

$$\langle f,\phi_n\rangle=\int_0^1 f(t)\overline{\phi_n(t)}\,dt$$

on $L^2[0,1]$. (See [1, pp. 234, 243].)

* Received by the editors September 1, 1981.
† Department of Mathematics, The University, Manchester M13 9PL, England.

If R is the operator with kernel $\Sigma_1^N \lambda_n \phi_n(x)\overline{\phi_n(t)}$, then R is the best approximation to T in the operator norm by symmetric operators of rank $\leq N$, the minimum distance being

$$\|R-T\|=|\lambda_{N+1}|.$$

Also $\Sigma_1^N \lambda_n \phi_n(x)\overline{\phi_n(t)}$ is the best approximation to $K(x,t)$ in mean square by symmetric kernels of rank $\leq N$, the minimum distance being $\Sigma_{N+1}^{\infty} \lambda_n^2$. (See [1, p. 239, 243].)

If we now assume $K(x,t)$ is positive definite and continuous, then we have Mercer's theorem, which says that the above eigenfunction expansions are uniformly absolutely convergent, also that $\lambda_n \geq 0$ for all n, $K(x,x)\geq 0$ for all x and

$$\sum_1^{\infty} \lambda_n = \int_0^1 K(x,x)\,dx < \infty.$$

(See [1, p. 245].) Kernels of this type therefore give rise to trace class operators, i.e., operators whose trace norm $\|T\|_{\mathrm{tr}} = \Sigma_1^{\infty} |\lambda_n|$ is finite.

LEMMA 1. *If $K(x,t)$ is symmetric and trace class, then $\Sigma_1^N \lambda_n \phi_n(x)\overline{\phi_n(t)}$ is the best approximation to $K(x,t)$ in the trace norm by symmetric kernels of rank $\leq N$.*

Proof. Let R be any symmetric operator of rank $\leq N$, and suppose that $T-R$ has eigenfunction expansion $\Sigma_1^{\infty} \mu_n \psi_n(x)\overline{\psi_n(t)}$. Then

$$S = R + \sum_1^p \mu_n \psi_n(x)\overline{\psi_n(t)}$$

has rank $\leq N+p$ and

$$T-S = \sum_{p+1}^{\infty} \mu_n \psi_n(x)\overline{\psi_n(t)}.$$

Therefore

$$|\mu_{p+1}| = \|T-S\| \geq |\lambda_{N+p+1}|.$$

Hence

$$\sum_{N+1}^{\infty} |\lambda_n| \leq \sum_1^{\infty} |\mu_n| = \|T-R\|_{\mathrm{tr}}.$$

However, for the operator R having kernel $\Sigma_1^N \lambda_n \phi_n(x)\overline{\phi_n(t)}$, we have

$$\|T-R\|_{\mathrm{tr}} = \sum_{N+1}^{\infty} |\lambda_n|,$$

so the lemma is proved. □

3. Square roots. Any positive trace class operator T has a unique positive square root S. In fact, if T has kernel $K(x,t)$ with eigenfunction expansion

$$K(x,t) = \sum_1^{\infty} \lambda_n \phi_n(x)\overline{\phi_n(t)},$$

then S must have kernel $J(x,t)$ given by

$$J(x,t) = \sum_1^{\infty} \lambda_n^{1/2} \phi_n(x)\overline{\phi_n(t)}.$$

LEMMA 2. *If $K(x,t)$ is continuous and positive definite and $J(x,t)$ is its positive square root, then, for any $f\in L^2[0,1]$,*

$$\int_0^1 J(x,t)f(t)\,dt$$

is a continuous function of x.

Proof.

$$\int_0^1 J(x,t)f(t)\,dt=\sum_1^\infty \lambda_n^{1/2}\langle f,\phi_n\rangle\phi_n(x)$$

and the series is uniformly absolutely convergent since

$$\sum_M^N |\lambda_n^{1/2}\langle f,\phi_n\rangle\phi_n(x)|\le\left(\sum_M^N \lambda_n|\phi_n(x)|^2\sum_M^N|\langle f,\phi_n\rangle|^2\right)^{1/2},$$

by the Cauchy–Schwarz inequality, and

$$\sum_1^\infty \lambda_n|\phi_n(x)|^2=K(x,x)$$

uniformly absolutely, whilst

$$\sum_M^N|\langle f,\phi_n\rangle|^2\le\|f\|^2$$

by Bessel's inequality. Every $\phi_n(x)$ is continuous, since $K(x,t)$ is, so the lemma follows. □

LEMMA 3. *If T is positive with continuous kernel and S is its positive square root and, if R is of finite rank and satisfies*

$$0\le R\le I,$$

where I is the identity operator, then SRS has continuous kernel and

$$0\le SRS\le T.$$

Proof. If R,S,T have kernels $H(x,t)$, $J(x,t)$, $K(x,t)$ respectively, then SRS has kernel

$$G(x,t)=\int_0^1\int_0^1 J(x,u)H(u,v)J(v,t)\,du\,dv.$$

Now $H(u,v)$ takes the form

$$H(u,v)=\sum_{i,j=1}^n a_{ij}\psi_i(u)\overline{\psi_j(v)},$$

where $a_{ji}=\overline{a_{ij}}$; therefore

$$G(x,t)=\sum_{i,j=1}^n a_{ij}\int_0^1 J(x,u)\psi_i(u)\,du\,\overline{\int_0^1 J(t,v)\psi_j(v)\,dv},$$

which is continuous by Lemma 2. The rest of the proof is easy. □

4. Proof of the result. Suppose $K(x,t)$ is positive definite and continuously differentiable. Then, for any $\varepsilon>0$, we can choose N such that

$$\left|\frac{\partial K}{\partial x}(x,t)-\frac{\partial K}{\partial x}(y,u)\right|<\varepsilon,\qquad \left|\frac{\partial K}{\partial t}(x,t)-\frac{\partial K}{\partial t}(y,u)\right|<\varepsilon$$

for all $x,y,t,u\in[0,1]$ satisfying

$$|x-y|<\frac{1}{N}, \qquad |t-u|<\frac{1}{N}.$$

Let R be the operator with kernel

$$H(x,t)=N\sum_1^N \psi_n(x)\psi_n(t),$$

where

$$\psi_n(x)=\begin{cases}1 & \text{if } \dfrac{n-1}{N}\le x\le\dfrac{n}{N},\\ 0 & \text{otherwise.}\end{cases}$$

Then

$$0\le R\le I$$

clearly, and so, by Lemma 3, SRS has a continuous kernel and

$$0\le SRS\le T,$$

where S is the positive square root of T. Therefore, by Mercer's theorem, we have

$$\begin{aligned}\|T-SRS\|_{\mathrm{tr}}&=\int_0^1 K(x,x)\,dx-\int_0^1\int_0^1\int_0^1 J(x,u)H(u,v)J(v,x)\,du\,dv\,dx\\&=\int_0^1 K(x,x)\,dx-\int_0^1\int_0^1 H(u,v)\left(\int_0^1 J(x,u)J(v,x)\,dx\right)du\,dv\\&=\int_0^1 K(x,x)\,dx-\int_0^1\int_0^1 H(u,v)K(v,u)\,du\,dv.\end{aligned}$$

Now

$$\int_0^1 K(x,x)\,dx=\int_0^1\int_0^1 H(u,v)K(u,u)\,du\,dv=\int_0^1\int_0^1 H(u,v)K(v,v)\,du\,dv,$$

since

$$\int_0^1 H(u,v)\,du=\int_0^1 H(u,v)\,dv=1$$

for all u,v. Therefore

$$\begin{aligned}\|T-SRS\|_{\mathrm{tr}}&=\int_0^1\int_0^1 H(u,v)\left[\frac{1}{2}(K(u,u)+K(v,v))-K(v,u)\right]du\,dv\\&=N\sum_1^N\int_{(n-1)/N}^{n/N}\int_{(n-1)/N}^{n/N}\left[\frac{1}{2}(K(u,u)+K(v,v))-K(v,u)\right]du\,dv.\end{aligned}$$

Note that

$$\frac{1}{2}(K(u,u)+K(v,v))-K(v,u)\ge 0$$

since

$$|K(u,v)|\le(K(u,u)K(v,v))^{1/2},$$

on account of the positive definiteness of $K(u,v)$,

$$\le\frac{1}{2}(K(u,u)+K(v,v)),$$

by the arithmetic geometric mean inequality. Also, if $|u-v|<1/N$, we have

$$0\leq\frac{1}{2}(K(u,u)+K(v,v))-K(v,u)$$
$$=\frac{1}{2}(u-v)\left[\frac{\partial K}{\partial u}(\xi,u)-\frac{\partial K}{\partial v}(v,\eta)\right],$$

for some ξ,η lying between u and v,

$$=\frac{1}{2}(u-v)\left[\left(\frac{\partial K}{\partial u}(c,c)+\varepsilon_1\right)-\left(\frac{\partial K}{\partial v}(c,c)+\varepsilon_2\right)\right],$$

where $c=\frac{1}{2}(u+v)$ and $|\varepsilon_1|<\varepsilon$, $|\varepsilon_2|<\varepsilon$,

$$=\frac{1}{2}(u-v)(\varepsilon_1-\varepsilon_2),$$

since $K(u,v)$ is symmetric,

$$<\frac{\varepsilon}{N}.$$

If $K(u,v)$ is complex it is only necessary to consider its real part since the contribution to the integral

$$\int_0^1\int_0^1 H(u,v)K(v,u)\,du\,dv$$

from the imaginary part of $K(v,u)$ is zero. Therefore

$$\|T-SRS\|_{\mathrm{tr}}\leq N\sum_1^N\int_{(n-1)/N}^{n/N}\int_{(n-1)/N}^{n/N}\frac{\varepsilon}{N}\,du\,dv=\frac{\varepsilon}{N}.$$

But R, and therefore SRS, has rank N, so, by Lemma 1, we have

$$\sum_{N+1}^{\infty}\lambda_n\leq\frac{\varepsilon}{N}.$$

Hence

$$\sum_{N+1}^{\infty}\lambda_n=o\left(\frac{1}{N}\right)$$

as $N\to\infty$, from which it follows that

$$\lambda_n=o\left(\frac{1}{n^2}\right)$$

as $n\to\infty$.

5. Extensions and generalisations. If $K(x,t)$ is Lip_1 (and positive definite) then the same argument proves

$$\lambda_n=O\left(\frac{1}{n^2}\right)$$

as $n\to\infty$.

These results are best possible since e.g. $\Sigma_1^\infty(\cos 2\pi n(x-t)/n^2)$ is Lip_1, whilst $\Sigma_1^\infty(\cos 2\pi n(x-t)/n^\alpha)$ is C^1 for any $\alpha>2$.

Both results clearly extend to the case where $K(x,t)$ has a finite number of negative eigenvalues. This is a consequence of the fact that, if $K(x,t)$ is C^1, then the

eigenfunctions are C^1, and so the positive part of $K(x,t)$ is C^1. Similar considerations apply if $K(x,t)$ is Lip_1.

In [2], Weyl also showed that, if $K(x,t)$ is C^p, then

$$\lambda_n = o\left(\frac{1}{n^{p+1/2}}\right)$$

as $n \to \infty$. One might conjecture that, if also $K(x,t)$ is positive definite, then

$$\lambda_n = o\left(\frac{1}{n^{p+1}}\right)$$

as $n \to \infty$. Unfortunately, we have so far been unable to prove this.

REFERENCES

[1] F. Riesz and B. S. Nagy, *Functional Analysis*, Ungar, New York, 1952.

[2] H. Weyl, *Das Asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen*, Math. Ann., 71 (1912), pp. 441–479.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0013 $01.25/0

PERIODICALLY INVARIANT LINEAR SYSTEMS*

KLAUS BARBEY†

Abstract. The topic of this paper is the analysis of abstract linear systems on a locally compact group G, that is, of continuous linear operators $N: \mathcal{D}(G) \to \mathcal{D}'(G)$. Here, $\mathcal{D}(G)$ denotes the space of test functions on G with its inductive limit topology as introduced by Bruhat, Maurin and Kac and $\mathcal{D}'(G)$ the space of distributions on G with the weak topology. We call such a linear system N periodically invariant with respect to a given closed subgroup Γ, if N commutes with translations from Γ. Periodically invariant systems are of interest, e.g., in the theory of electrical networks with periodically varying parameters or in process control theory. Under the assumption that the quotient group G/Γ is compact, a Fourier series representation for Γ-periodic distributions on G is derived. From this we conclude via the Schwartz kernel theorem that the classical convolution representation for a translation invariant system ($\Gamma = G$) generalizes to a Fourier superposition $\Sigma \chi N_\chi$ of translation invariant systems N_χ, where summation runs over all characters $\chi \in \hat{G}$ vanishing on Γ. Finally, it is proved that N is causal with respect to a semigroup $P \subset G$ if and only if all individual systems N_χ are causal.

1. Introduction. The realizability theory of abstract linear systems as developed in Zemanian's book [17] studies continuous linear mappings N from the space $\mathcal{D}(G)$ of Schwartz test functions on $G = \mathbf{R}$ or $G = \mathbf{R}^n$ into the space of distributions $\mathcal{D}'(G)$ with the weak topology. If N is translation invariant, that is, if

$$N(\tau_u \varphi) = \tau_u N(\varphi) \quad \forall u \in G, \quad \forall \varphi \in \mathcal{D}(G),$$

where τ_u denotes the translation operator $\tau_u \varphi(t) = \varphi(t-u)$ for all $t \in G$, then it is well known that N can be realized as convolution with a fixed distribution $V \in \mathcal{D}'(G)$:

$$N\varphi = V * \varphi \quad \forall \varphi \in \mathcal{D}(G).$$

Under additional conditions such as causality and dissipativity a more detailed representation of N is possible [10], [7].

In the present paper we consider a situation which is more general in two respects. First, following, e.g., Fourès and Segal [5], Falb and Freedman [4], Hackenbroch [8] and Bose [2], G is allowed to be any locally compact abelian group. The Schwartz space of test functions is then replaced by its natural generalization introduced by Bruhat [3], Maurin [11] and Kac [9]. Second, the assumption of translation invariance is weakened: we suppose merely that the system N is periodically invariant, that is,

$$N(\tau_u \varphi) = \tau_u N(\varphi) \quad \forall u \in \Gamma, \quad \forall \varphi \in \mathcal{D}(G),$$

where $\Gamma \subset G$ is a given (without loss of generality closed) subgroup which is not too small in the sense that G/Γ is compact. It is then shown that the above convolution representation $N\varphi = V * \varphi$ for a translation invariant system N generalizes to a Fourier superposition

$$N\varphi = \sum_{\chi \in \Gamma^\perp} \chi (V_\chi * \varphi) \quad \text{with } V_\chi \in \mathcal{D}'(G) \quad \forall \chi \in \Gamma^\perp = \{\chi \in \hat{G} : \chi(\Gamma) = 1\}$$

in case that N is only periodically invariant. If $G = \mathbf{R}$ this result is due to Willie [16], if $G = \mathbf{R}^n$ it is found in [1]. Starting from electrical network and process control theory, periodically invariant linear systems were considered by Unbehauen [14], [15], where also a forerunner of the above formula can be found.

*Received by the editors June 22, 1981.
†Fakultät für Mathematik, Universität Regensburg, D 8400 Regensburg, West Germany.

The paper is organized as follows. Section 2 recalls in brief the definition of $\mathcal{D}(G)$ for a locally compact group G as given by Bruhat. For all details and proofs, the reader is referred to the original paper [3]. In §3 a Fourier series representation is established for Γ-periodic distributions on G which yields in §4 the above Fourier series realization for a Γ-invariant system N via the Schwartz kernel theorem. Finally it is shown in §5 that N is causal with respect to a given semigroup P if and only if all systems $\varphi \mapsto V_\chi * \varphi$ for all $\varphi \in \mathcal{D}(G)$ in the decomposition of N are causal with respect to P which in turn corresponds to carrier conditions on the distributions $V_\chi \in \mathcal{D}'(G)$.

2. Distributions on locally compact groups. In this section we provide some definitions and facts from the theory of distributions on a locally compact group G as developed in [3]. In case of a Lie group G (or more generally in case of a differentiable manifold G), $\mathcal{D}(G)$ is the usual space of Schwartz test functions with its inductive limit topology. In case of an arbitrary locally compact group G, the definition of $\mathcal{D}(G)$ is given in two steps: Let $\mathfrak{G}$ be the set of all compact invariant subgroups $k \subset G$ such that G/k is a Lie group; $\mathfrak{G}$ is directed downwards by inclusion. For $k \in \mathfrak{G}$ we denote by π_k the canonical mapping $G \to G/k$, and by $\mathcal{D}^k(G)$ the complex vector space $\{\psi \circ \pi_k : \psi \in \mathcal{D}(G/k)\}$ equipped with the topology inherited from $\mathcal{D}(G/k)$. Hence, a function $\varphi: G \to \mathbf{C}$ belongs to $\mathcal{D}^k(G)$ if and only if it is constant on the left cosets modulo k and, regarded then as a function $\tilde{\varphi}: G/k \to \mathbf{C}$, belongs to $\mathcal{D}(G/k)$. If G satisfies $\bigcap_{k \in \mathfrak{G}} k = \{e\}$, then G is called Lie projective and the space $\mathcal{D}(G)$ is defined to be the inductive limit of the spaces $\mathcal{D}^k(G)$ $(k \in \mathfrak{G})$. It is known [12, p. 175] that G is Lie projective if G/G_0 is compact, where G_0 is the component of the identity element $e \in G$. In general, G contains an open Lie projective subgroup G_1 [12, p. 54]. Thus, $\mathcal{D}(G_1)$ is defined, and for a left coset $\Delta = xG_1$, let $\mathcal{D}(\Delta)$ be the space of all functions $t \mapsto \varphi(x^{-1}t)$ for all $t \in \Delta$, where $\varphi \in \mathcal{D}(G_1)$, together with the topology induced from $\mathcal{D}(G_1)$. Then $\mathcal{D}(\Delta)$ is independent from $x \in \Delta$ and $\mathcal{D}(G)$ is defined to be the locally convex direct sum of the spaces $\mathcal{D}(\Delta)$, where Δ varies over all left cosets modulo G_1. Furthermore, $\mathcal{E}(G)$ is the space of all functions $\psi: G \to \mathbf{C}$ such that $\varphi\psi \in \mathcal{D}(G)$ for all $\varphi \in \mathcal{D}(G)$ together with the coarsest locally convex topology that makes continuous all mappings $\mathcal{E}(G) \to \mathcal{D}(G)$ defined by $\psi \mapsto \varphi\psi$ for all $\psi \in \mathcal{E}(G)$, where $\varphi \in \mathcal{D}(G)$. As usual, the elements of the dual space $\mathcal{D}'(G)$ (respective $\mathcal{E}'(G)$) are called distributions (respective distributions with compact support) on G. If $K \subset G$ is compact, the subspace $\mathcal{D}_K(G)$ consists of all $\varphi \in \mathcal{D}(G)$ with support $\operatorname{supp} \varphi \subset K$.

For a closed subgroup $\Gamma \subset G$, let m_Γ denote a left Haar measure on Γ regarded as a measure on G and normalized if Γ is compact. If a left Haar measure m_G on G is fixed, every continuous function $f: G \to \mathbf{C}$ gives rise to a distribution $[f]$ via $\varphi \mapsto \int \varphi f \, dm_G$ for all $\varphi \in \mathcal{D}(G)$; in this sense we have $\mathcal{C}(G) := \{f: G \to \mathbf{C} \text{ continuous}\} \subset \mathcal{D}'(G)$.

3. Γ-periodic distributions. Let G be a locally compact abelian group and $\Gamma \subset G$ a fixed closed subgroup. π is the canonical mapping from G onto the group G/Γ of left cosets. We call a distribution $V \in \mathcal{D}'(G)$ Γ-periodic if

$$V(\tau_u \varphi) = V(\varphi) \quad \forall u \in \Gamma, \quad \forall \varphi \in \mathcal{D}(G),$$

where τ_u denotes translation by $u \in G$, that is, $\tau_u \varphi(t) = \varphi(u^{-1}t)$ for all $t \in G$. In this section we establish a Fourier series expansion for Γ-periodic distributions in the case that G/Γ is compact. We start with the following remark which allows us to identify

Γ-periodic distributions with distributions on the group G/Γ. Let ν denote the continuous linear mapping from $\mathcal{D}(G)$ onto $\mathcal{D}(G/\Gamma)$ [3, p. 59] defined by

$$\nu(\varphi)(x\Gamma)=\int\varphi(xt^{-1})\,dm_\Gamma(t)=\varphi*m_\Gamma(x)\quad\forall x\in G,\quad\forall\varphi\in\mathcal{D}(G).$$

Remark 1. The mapping $U\mapsto V$: $V(\varphi)=U(\nu(\varphi))$ for all $\varphi\in\mathcal{D}(G)$ is a bijection between the distributions $U\in\mathcal{D}'(G/\Gamma)$ and the Γ-periodic distributions $V\in\mathcal{D}'(G)$; its inverse is given by $V\mapsto U$: $U(\psi)=V(\beta(\psi\circ\pi))$ for all $\psi\in\mathcal{D}(G/\Gamma)$ where β is any function $\beta\in\mathcal{E}(G)$ with $\operatorname{supp}\beta\cap\pi^{-1}(K)$ compact for $K\subset G/\Gamma$ compact and $\int\beta(xt^{-1})\,dm_\Gamma(t)=1$ for all $x\in G$.

The existence of such functions $\beta\in\mathcal{E}(G)$ is proved in [3, p. 59]. For a Γ-periodic $V\in\mathcal{D}'(G)$, Remark 1 shows that $V(\beta(\psi\circ\pi))$ ($\psi\in\mathcal{D}(G/\Gamma)$) is independent of the particular $\beta\in\mathcal{E}(G)$.

Proof. 1) Since $\psi\mapsto\beta(\psi\circ\pi)$ is a continuous linear mapping $\mathcal{D}(G/\Gamma)\to\mathcal{D}(G)$, the mapping $V\mapsto U$: $U(\psi)=V(\beta(\psi\circ\pi))$ for all $\psi\in\mathcal{D}(G/\Gamma)$ is well defined, furthermore, it satisfies

$$V(\beta(\psi\circ\pi))=U(\nu(\beta(\psi\circ\pi)))=U(\psi)\quad\forall\psi\in\mathcal{D}(G/\Gamma)$$

if $U\in\mathcal{D}'(G/\Gamma)$ and $V=U\circ\nu\in\mathcal{D}'(G)$.

2) Now let $V\in\mathcal{D}'(G)$ be Γ-periodic and define $U\in\mathcal{D}'(G/\Gamma)$ by $U(\psi)=V(\beta(\psi\circ\pi))$ for all $\psi\in\mathcal{D}(G/\Gamma)$. We have to show that $U\circ\nu=V$. To this end fix some $\varphi\in\mathcal{D}(G)$ and choose a function $\alpha\in\mathcal{D}(G)$ with $\alpha=1$ on $\operatorname{supp}\beta\cap\pi^{-1}(\pi(\operatorname{supp}\varphi))$. Define $S,T\in\mathcal{E}'(G)$ by

$$S(\psi)=V(\varphi\psi)\quad\text{and}\quad T(\psi)=V(\alpha\beta\psi)\quad\forall\psi\in\mathcal{E}(G).$$

Then the Γ-periodicity of V implies that

$$\begin{aligned}U(\nu(\varphi))&=V(\beta(\varphi*m_\Gamma))=V(\alpha\beta(\varphi*m_\Gamma))=T(\varphi*m_\Gamma)\\&=T*m_\Gamma(\varphi)=\int T(\varphi(\cdot\,t^{-1}))dm_\Gamma(t)=\int V(\alpha\beta\varphi(\cdot\,t^{-1}))dm_\Gamma(t)\\&=\int V(\alpha(\cdot\,t)\beta(\cdot\,t)\varphi)\,dm_\Gamma(t)=\int S((\alpha\beta)(\cdot\,t^{-1}))\,dm_\Gamma(t)=S*m_\Gamma(\alpha\beta)\\&=S((\alpha\beta)*m_\Gamma)=V\Big(\varphi\int\alpha(\cdot\,t^{-1})\beta(\cdot\,t^{-1})\,dm_\Gamma(t)\Big)=V(\varphi).\end{aligned}$$

Q.E.D.

It is a remarkable feature of Bruhat's theory that the coefficients of finite dimensional representations of G belong to $\mathcal{E}(G)$ [3, p. 50], in particular, the dual group $\hat{G}$ satisfies $\hat{G}\subset\mathcal{E}(G)$. Hence, for a distribution U on a compact abelian group G, we can define the Fourier coefficients $\hat{U}(\chi)\in\mathbf{C}$ ($\chi\in\hat{G}$) by

$$\hat{U}(\chi)=U(\bar{\chi})\quad\forall\chi\in\hat{G},$$

where the bar $^-$ denotes complex conjugation.

Remark 2. Assume that G is compact and abelian. Then every distribution $U\in\mathcal{D}'(G)$ has the unique Fourier series expansion

$$U=\sum_{\chi\in\hat{G}}\hat{U}(\chi)[\chi];$$

the series is weakly summable, that means $U(\varphi)=\Sigma_{\chi\in\hat{G}}\hat{U}(\chi)\int\varphi\chi\,dm_G$ for all $\varphi\in\mathcal{D}(G)$, where the series is summable (= the partial sums over finite subsets of $\hat{G}$ converge to $U(\varphi)$).

Proof. The uniqueness of the representation is clear from the orthogonality of the characters $\chi \in \hat{G}$. Hence, it suffices to show that every $\varphi \in \mathcal{D}(G)$ has the representation $\varphi = \Sigma_{\chi \in \hat{G}} \hat{\varphi}(\chi)\chi$, where the series is summable in the topology of $\mathcal{D}(G)$ and $\hat{\varphi}(\chi) = \int \varphi \bar{\chi}\, dm_G$ for all $\chi \in \hat{G}$. This is easily deduced from classical facts in case that G is a Lie group since then $G = T \times H$, with T a torus group and H a finite abelian group. Now, since any compact group is Lie projective, for every $\varphi \in \mathcal{D}(G)$ there is a $k \in \mathfrak{G}$ such that φ is invariant under k with corresponding function $\tilde{\varphi} \in \mathcal{D}(G/k)$, hence $\tilde{\varphi} = \Sigma_{\tilde{\chi} \in (G/k)^{\wedge}} \hat{\tilde{\varphi}}(\tilde{\chi})\tilde{\chi}$ where the series is summable in $\mathcal{D}(G/k)$. In view of $(G/k)^{\wedge} \equiv k^{\perp} = \{\chi \in \hat{G}: \chi = 1 \text{ on } k\}$ via the identification $\tilde{\chi} \equiv \chi$, this implies $\varphi = \Sigma_{\chi \in k^{\perp}} \hat{\tilde{\varphi}}(\tilde{\chi})\chi$, where the series is summable in $\mathcal{D}^k(G)$. Finally, we obtain

$$\hat{\varphi}(\chi) = \int_{G/k} \tilde{\varphi}(xk) \left(\int_k \chi(xt)\, dm_k(t) \right) dm_{G/k}(xk) = \begin{cases} \hat{\tilde{\varphi}}(\tilde{\chi}) & \text{if } \chi \in k^{\perp}, \\ 0 & \text{if } \chi \notin k^{\perp}. \end{cases}$$ Q.E.D.

Now we come to the main result of this section. If G/Γ is compact we define for a Γ-periodic distribution $V \in \mathcal{D}'(G)$ the Fourier coefficients $\hat{V}(\chi) \in \mathbf{C}(\chi \in \Gamma^{\perp})$ of V by

$$\hat{V}(\chi) = \hat{U}(\tilde{\chi}) = V(\beta\bar{\chi}) \quad \forall \chi \in \Gamma^{\perp} \equiv (G/\Gamma)^{\wedge}$$

where U is the distribution on G/Γ corresponding to V and β is any function as in Remark 1. Observe that the compactness of G/Γ implies that even $\beta \in \mathcal{D}(G)$.

THEOREM 1. *Let G be a locally compact abelian group and $\Gamma \subset G$ a closed subgroup such that G/Γ is compact. Then every Γ-periodic distribution $V \in \mathcal{D}'(G)$ admits a Fourier series representation*

$$V = \sum_{\chi \in \Gamma^{\perp}} \hat{V}(\chi)[\chi],$$

where the series is weakly summable.

Proof. Let $\varphi \in \mathcal{D}(G)$ and let U denote the distribution on G/Γ corresponding to V. We apply Remark 2 with G replaced by G/Γ to obtain

$$V(\varphi) = U(\nu(\varphi)) = \sum_{\tilde{\chi} \in (G/\Gamma)^{\wedge}} \hat{U}(\tilde{\chi}) \int_{G/\Gamma} \nu(\varphi)\tilde{\chi}\, dm_{G/\Gamma}$$

and hence the result since

$$\int_G \varphi\chi\, dm_G = \int_{G/\Gamma} \left(\int_{\Gamma} \varphi(xt)\chi(xt)\, dm_{\Gamma}(t) \right) dm_{G/\Gamma}(x\Gamma) = \int_{G/\Gamma} \nu(\varphi)\tilde{\chi}\, dm_{G/\Gamma}$$

for all $\chi \in \Gamma^{\perp} \equiv (G/\Gamma)^{\wedge}$. It remains to prove the uniqueness assertion. Assume that $V = \Sigma_{\chi \in \Gamma^{\perp}} c_{\chi}[\chi]$ with any $c_{\chi} \in \mathbf{C}$ and fix $\eta \in \Gamma^{\perp}$. The orthogonality of the characters then implies that

$$\begin{aligned} \hat{V}(\eta) = V(\beta\bar{\eta}) &= \sum_{\chi \in \Gamma^{\perp}} c_{\chi} \int \beta\chi\bar{\eta}\, dm_G \\ &= \sum_{\chi \in \Gamma^{\perp}} c_{\chi} \int_{G/\Gamma} \left(\int_{\Gamma} \beta(xt)\chi(xt)\overline{\eta(xt)}\, dm_{\Gamma}(t) \right) dm_{G/\Gamma}(x\Gamma) \\ &= \sum_{\chi \in \Gamma^{\perp}} c_{\chi} \int_{G/\Gamma} \tilde{\chi}(x\Gamma)\overline{\tilde{\eta}(x\Gamma)}\, dm_{G/\Gamma}(x\Gamma) = c_{\eta}. \end{aligned}$$ Q.E.D.

4. Periodically invariant linear systems. Let G be a locally compact abelian group and $N\colon \mathcal{D}(G)\to\mathcal{D}'(G)$ a linear system, that is a linear operator which is continuous with respect to the weak topology on $\mathcal{D}'(G)$. Suppose that $\Gamma\subset G$ is a subgroup. We call N Γ-invariant if

$$N(\tau_u\varphi)(\psi)=\tau_u N(\varphi)(\psi):=N(\varphi)(\tau_{u^{-1}}\psi)\quad\forall u\in\Gamma,\quad\forall\varphi,\psi\in\mathcal{D}(G).$$

Since the translation $u\mapsto\tau_u\varphi$ is continuous on G for every $\varphi\in\mathcal{D}(G)$, we may assume Γ to be closed. The most important example is of course $G=\mathbf{R}^n$ and $\Gamma=\{(k_1c_1,\cdots,k_nc_n)\colon k_1,\cdots,k_n=0,\pm1,\pm2,\cdots\}\subset\mathbf{R}^n$, with fixed period length $c_1,\cdots,c_n>0$. The aim of this section is to prove the following theorem.

THEOREM 2. *Assume that G is a locally compact abelian group and $\Gamma\subset G$ is a closed subgroup such that G/Γ is compact. Let $N\colon\mathcal{D}(G)\to\mathcal{D}'(G)$ be a Γ-invariant linear system. Then for every character $\chi\in\Gamma^\perp$ there exists a unique distribution $V_\chi\in\mathcal{D}'(G)$ such that*

$$N\varphi=\sum_{\chi\in\Gamma^\perp}\chi\big[V_\chi*\varphi\big]\quad\forall\varphi\in\mathcal{D}(G),$$

where the series is weakly summable, that is,

$$N\varphi(\psi)=\sum_{\chi\in\Gamma^\perp}\int(V_\chi*\varphi)\chi\psi\,dm_G\quad\forall\varphi,\psi\in\mathcal{D}(G).$$

The proof of the theorem leans heavily on the Schwartz kernel theorem in the version of Bruhat [3, pp. 55–56]: For every linear system $N\colon\mathcal{D}(G)\to\mathcal{D}'(G)$, there exists a unique distribution $S\in\mathcal{D}'(G\times G)$ such that

$$N\varphi(\psi)=S(\varphi\otimes\psi)\quad\forall\varphi,\psi\in\mathcal{D}(G),$$

where $\varphi\otimes\psi\in\mathcal{D}(G\times G)$ denotes the function $\varphi\otimes\psi(s,t)=\varphi(s)\psi(t)$ for all $s,t\in G$. Furthermore, N is Γ-invariant if and only if S satisfies $S(\tau_{(u,u)}\Phi)=S(\Phi)$ for all $u\in\Gamma$ for all $\Phi\in\mathcal{D}(G\times G)$.

Another consequence of the kernel theorem is the following. For every two distributions U,V on G, there exists a unique distribution on $G\times G$ denoted by $U\otimes V$ which satisfies $U\otimes V(\varphi\otimes\psi)=U(\varphi)V(\psi)$ for all $\varphi,\psi\in\mathcal{D}(G)$. The subsequent remark follows easily from the definitions.

Remark 3. Assume that $S\in\mathcal{D}'(G\times G)$ satisfies $S(\tau_{(u,u)}\Phi)=S(\Phi)$ for all $u\in\Gamma$ for all $\Phi\in\mathcal{D}(G\times G)$ and define $C\colon G\times G\to G\times G$ by $C(s,t)=(s^{-1}t,t)$ for all $s,t\in G$. Then $C\circ C$ is the identity function and $W\colon W(\Phi)=S(\Phi\circ C)$ for all $\Phi\in\mathcal{D}(G\times G)$ is a distribution on $G\times G$ such that $W(\tau_{(e,u)}\Phi)=W(\Phi)$ for all $u\in\Gamma$ and $S(\Phi)=W(\Phi\circ C)$ for all $\Phi\in\mathcal{D}(G\times G)$.

Proof of Theorem 2. The kernel theorem via the above remark supplies us with a distribution $W\in\mathcal{D}'(G\times G)$ such that $W(\tau_{(e,u)}\Phi)=W(\Phi)$ for all $u\in\Gamma$ for all $\Phi\in\mathcal{D}(G\times G)$. For fixed $\varphi\in\mathcal{D}(G)$, we set $W_\varphi(\psi)=W(\varphi\otimes\psi)$ for all $\psi\in\mathcal{D}(G)$. Then $W_\varphi\in\mathcal{D}'(G)$ is Γ-periodic; hence, Theorem 1 implies that

$$W(\varphi\otimes\psi)=W_\varphi(\psi)=\sum_{\chi\in\Gamma^\perp}V_\chi(\varphi)[\chi](\psi)=\sum_{\chi\in\Gamma^\perp}V_\chi\otimes[\chi](\varphi\otimes\psi)\quad\forall\varphi,\psi\in\mathcal{D}(G),$$

where $V_\chi\in\mathcal{D}'(G)$ is defined by

$$V_\chi(\varphi)=\hat{W}_\varphi(\chi)=W_\varphi(\beta\bar{\chi})=W(\varphi\otimes\beta\bar{\chi})\quad\forall\varphi\in\mathcal{D}(G).$$

Now we apply Lemma 2 to obtain

$$W(\Phi)=\sum_{\chi\in\Gamma^{\perp}} V_{\chi}\otimes[\chi](\Phi),$$

$$S(\Phi)=\sum_{\chi\in\Gamma^{\perp}} V_{\chi}\otimes[\chi](\Phi\circ C),\quad \forall\Phi\in\mathcal{D}(G\times G).$$

For $V\in\mathcal{D}'(G)$ define $\check{V}\in\mathcal{D}'(G)$ by $\check{V}(\varphi)=V(\check{\varphi})$ and $\check{\varphi}(t)=\varphi(t^{-1})$ for all $\varphi\in\mathcal{D}(G)$ for all $t\in G$. Then an easy calculation shows that

$$V_{\chi}\otimes[\chi]((\varphi\otimes\psi)\circ C)=\check{V}_{\chi}*[\psi\chi](\varphi)=[\psi\chi](V_{\chi}*\varphi)\quad\forall\varphi,\psi\in\mathcal{D}(G),$$

and, hence,

$$N\varphi(\psi)=S(\varphi\otimes\psi)=\sum_{\chi\in\Gamma^{\perp}}[\psi\chi](V_{\chi}*\varphi)=\sum_{\chi\in\Gamma^{\perp}}\int(V_{\chi}*\varphi)\chi\psi\,dm_{G}\quad\forall\varphi,\psi\in\mathcal{D}(G).$$

It remains to prove the uniqueness assertion. Suppose that $U_{\chi}\in\mathcal{D}'(G)$ for all $\chi\in\Gamma^{\perp}$ such that

$$S(\varphi\otimes\psi)=\sum_{\chi\in\Gamma^{\perp}}\int(U_{\chi}*\varphi)\chi\psi\,dm_{G}=\sum_{\chi\in\Gamma^{\perp}}U_{\chi}\otimes[\chi]((\varphi\otimes\psi)\circ C)\quad\forall\varphi,\psi\in\mathcal{D}(G).$$

We apply Lemma 2 again with S_{χ}: $S_{\chi}(\Phi)=U_{\chi}\otimes[\chi](\Phi\circ C)$ for all $\Phi\in\mathcal{D}(G\times G)$ to obtain

$$W_{\varphi}(\psi)=W(\varphi\otimes\psi)=S((\varphi\otimes\psi)\circ C)=\sum_{\chi\in\Gamma^{\perp}}U_{\chi}\otimes[\chi](\varphi\otimes\psi)$$

$$=\sum_{\chi\in\Gamma^{\perp}}U_{\chi}(\varphi)\int\chi\psi\,dm_{G}\quad\forall\varphi,\psi\in\mathcal{D}(G).$$

So we conclude $U_{\chi}(\varphi)=\hat{W}_{\varphi}(\chi)=V_{\chi}(\varphi)$ for all $\chi\in\Gamma^{\perp}$ from the uniqueness assertion in Theorem 1. Q.E.D.

Now we must prove the subsequent Lemma 2 which we have used in the proof of the theorem. We shall need the following fact, which may be well known; in any case it admits a standard proof via the Banach–Steinhaus theorem.

LEMMA 1. *Let E,F be locally convex barrelled spaces. Assume that $(b_i)_{i\in I}$ is a family of separately continuous bilinear forms on $E\times F$ such that $\Sigma_{i\in I}b_i(u,v)$ is summable for all $u\in F$. Then the convergence of $\Sigma_{i\in I}b_i$ is uniform on products $P\times Q$ of precompact subsets $P\subset E$ and $Q\subset F$.*

Proof. 1) Suppose that $I=\{1,2,\cdots\}$ and that $\Sigma_{n=1}^{\infty}b_n(u,v)$ converges for all $u\in E$ for all $v\in F$. Then an iterated application of the Banach–Steinhaus theorem, e.g., in the versions of [13, p. 69], shows that the convergence is uniform on $P\times Q$.

2) Now let I be an arbitrary index set. If the assertion is assumed to be false, then there is a $\varepsilon>0$ and a sequence $\delta(1)$, $\delta(2),\cdots$ of disjoint finite subsets $\subset I$ such that $\sup_{u\in P,v\in Q}|\Sigma_{i\in\delta(n)}b_i(u,v)|>\varepsilon$ for all $n=1,2,\cdots$. Apply 1) with $a_n=\Sigma_{i\in\delta(n)}b_i$ ($n=1,2,\cdots$), to arrive at a contradiction. Q.E.D.

LEMMA 2. *Assume that $(S_i)_{i\in I}$ is a family of distributions on $G\times G$ such that $\Sigma_{i\in I}S_i(\varphi\otimes\psi)$ is summable for all $\varphi,\psi\in\mathcal{D}(G)$. Then $\Sigma_{i\in I}S_i(\Phi)$ is summable for all $\Phi\in\mathcal{D}(G\times G)$ and defines a distribution on $G\times G$.*

Proof. Without loss of generality, we may assume that G is Lie projective.

1) Fix some $\Phi\in\mathcal{D}(G\times G)$ and denote by K (respective L) the projection of $\operatorname{supp}\Phi$ onto the first (respective second) coordinate. There is a $k\in\mathfrak{G}$ such that $\Phi\in\mathcal{D}^{k\times k}(G\times G)$; hence the corresponding function $\tilde{\Phi}$ on $G/k\times G/k$ satisfies $\tilde{\Phi}\in\mathcal{D}_{\pi(K)\times\pi(L)}(G/k\times G/k)=\mathcal{D}_{\pi(K)}(G/k)\hat{\otimes}\mathcal{D}_{\pi(L)}(G/k)$, where the inductive and projective tensor products coincide since both factors are Fréchet spaces [6, Chap. II p. 84]. Now a theorem of Grothendieck [6, Chap. I p. 51] implies the existence of a sequence (λ_n) in $\mathbf{C}$ with $\Sigma_{n=1}^{\infty}|\lambda_n|\leq 1$ and of sequences $(\tilde{\varphi}_n)$ in $\mathcal{D}_{\pi(K)}(G/k)$, $(\tilde{\psi}_n)$ in $\mathcal{D}_{\pi(L)}(G/k)$, both converging to zero, such that $\tilde{\Phi}=\Sigma_{n=1}^{\infty}\lambda_n\tilde{\varphi}_n\otimes\tilde{\psi}_n$, hence $\Phi=\Sigma_{n=1}^{\infty}\lambda_n\varphi_n\otimes\psi_n$, with sequences (φ_n), (ψ_n) in $\mathcal{D}(G)$, both converging to zero. For later application we note that even $\varphi_n\in\mathcal{D}_K(G)$ and $\psi_n\in\mathcal{D}_L(G)$ for all $n=1,2,\cdots$ as follows from $\operatorname{supp}\Phi\subset(\operatorname{supp}\Phi)k\times k$.

2) Given $\varepsilon>0$, we obtain from Lemma 1 a finite subset $\alpha\subset I$ such that $|\Sigma_{i\in\delta}S_i(\varphi_n\otimes\psi_n)|\leq\varepsilon$ for all $n=1,2,\cdots$ for every finite subset $\delta\subset I$ disjoint with α. We conclude that

$$\left|\sum_{i\in\delta}S_i\left(\sum_{n=1}^{r}\lambda_n\varphi_n\otimes\psi_n\right)\right|\leq\sum_{n=1}^{r}|\lambda_n|\left|\sum_{i\in\delta}S_i(\varphi_n\otimes\psi_n)\right|\leq\varepsilon$$

and hence $|\Sigma_{i\in\delta}S_i(\Phi)|\leq\varepsilon$ for the $\delta\subset I$ in question.

3) Since $\Sigma_{i\in I}S_i(\Phi)$ is summable, it is absolutely summable, in particular, the family $(\Sigma_{i\in\alpha}S_i\colon\ \alpha\subset I \text{ finite})$ is pointwise bounded. This gives the continuity of $\Sigma_{i\in I}S_i$ via the Banach–Steinhaus theorem. Q.E.D.

The following corollary is (in case of a separable group) due to Bose [2].

COROLLARY. *Let G be a locally compact abelian group and assume that $N\colon\mathcal{D}(G)\to\mathcal{D}'(G)$ is a translation-invariant linear system. Then there exists a unique distribution $V\in\mathcal{D}'(G)$ such that*

$$N\varphi=[V*\varphi]\quad\forall\varphi\in\mathcal{D}(G).$$

Let us now consider a system which is regular into $\mathcal{C}(G)$, that means a linear system $N\colon\mathcal{D}(G)\to\mathcal{C}(G)\subset\mathcal{D}'(G)$. The closed graph theorem then shows that N is continuous even with respect to the usual topology on $\mathcal{C}(G)$. As in [16], we introduce a function $Z\colon G\to\mathcal{D}'(G)$ by

$$Z(x)\varphi=N(\tau_x\check{\varphi})(x)\quad\forall x\in G,\quad\forall\varphi\in\mathcal{D}(G).$$

Then Z is continuous with respect to the weak topology on $\mathcal{D}'(G)$ and represents N in the sense that

$$N\varphi(x)=(Z(x)*\varphi)(x)\quad\forall x\in G,\quad\forall\varphi\in\mathcal{D}(G);$$

furthermore, N is Γ-invariant if and only if the function Z is Γ-periodic, as follows easily from the definitions. In particular, N is translation invariant if and only if Z is constant on G. Assume again that G/Γ is compact. If N is Γ-invariant and if N is represented according to Theorem 2 as $N\varphi=\Sigma_{\chi\in\Gamma^{\perp}}\chi[V_\chi*\varphi]$, then we have $V_\chi=\hat{Z}(\chi)$, where the Fourier coefficients $\hat{Z}(\chi)\in\mathcal{D}'(G)$ $(\chi\in\Gamma^{\perp})$ are given by

$$\hat{Z}(\chi)\varphi=\int(\beta\bar{\chi})(x)Z(x)\varphi\,dm_G(x)\quad\forall\varphi\in\mathcal{D}(G),$$

with $\beta\in\mathcal{D}(G)$ as in §2.

5. Causality. In this section we study Γ-invariant linear systems on a locally compact abelian group G under the additional assumption of causality. We fix a closed semigroup $P \subset G$, that is, a subset $P \subset G$ such that $e \in P$ and $PP \subset P$; furthermore, we assume that $e \in \operatorname{int} P$, where $\operatorname{int} P$ means the topological interior of P. A linear system N: $\mathcal{D}(G) \to \mathcal{D}'(G)$ is called P-causal if $\varphi = 0$ on uP^{-1} implies $N\varphi = 0$ on $u(\operatorname{int} P)^{-1}$ for all $u \in G$ for all $\varphi \in \mathcal{D}(G)$. In the classical case of a translation invariant linear system $N = V*$, it is known that N is P-causal if and only if $\operatorname{supp} V \subset P$. Our goal is to establish the corresponding result in the Γ-invariant case.

LEMMA 3. *Assume that N: $\mathcal{D}(G) \to \mathcal{D}'(G)$ is a linear system and let S denote the distribution on $G \times G$ such that $N\varphi(\psi) = S(\varphi \otimes \psi)$ for all $\varphi, \psi \in \mathcal{D}(G)$. Then N is P-causal if and only if* $\operatorname{supp} S \subset \{(s,t)\colon s,t \in G,\ s^{-1}t \in P\}$.

Proof. 1) We set $Q = \{(s,t)\colon s,t \in G,\ s^{-1}t \notin P\}$ and $Q_u = \{(s,t)\colon s,t \in G,\ s \notin uP^{-1},\ t \in u(\operatorname{int} P)^{-1}\}$ for all $u \in G$. From our topological assumption on P it is easy to deduce that $Q = \bigcup_{u \in G} Q_u$.

2) $\Rightarrow$: Let G_1 be an open Lie projective subgroup of G. Then $G_1 \times G_1$ is an open Lie projective subgroup of $G \times G$. Consider a $\Phi \in \mathcal{D}(G \times G)$ such that $\operatorname{supp} \Phi \subset Q_u$ for some $u \in G$ and let $K \subset G \setminus uP^{-1}$ (respective $L \subset u(\operatorname{int} P)^{-1}$) denote the projection of $\operatorname{supp} \Phi$ onto the first (respective second) coordinate. By definition, Φ is a finite sum $\Phi = \sum_{l=1}^{r} \tau_{(x(l),y(l))} \Psi_l$ with $x(1), \cdots, x(r), y(1), \cdots, y(r) \in G$ and $\Psi_1, \cdots, \Psi_r \in \mathcal{D}(G_1 \times G_1)$. As in the proof of Lemma 2, we conclude that $\Psi_l = \sum_{n=1}^{\infty} \lambda_n^l \varphi_n^l \otimes \psi_n^l$ with $\sum_{n=1}^{\infty} |\lambda_n^l| \leq 1$ and sequences $(\varphi_n^l)_n$ in $\mathcal{D}_{x(l)^{-1}K}(G_1)$, $(\psi_n^l)_n$ in $\mathcal{D}_{y(l)^{-1}L}(G_1)$, both converging to zero ($l = 1, \cdots, r$). The P-causality of N then implies that

$$S(\Phi) = \sum_{l=1}^{r} \sum_{n=1}^{\infty} \lambda_n^l S\left(\tau_{x(l)} \varphi_n^l \otimes \tau_{y(l)} \psi_n^l\right) = 0.$$

Thus we have $S = 0$ on Q_u for all $u \in G$; hence from 1) we see that $S = 0$ on Q since $\mathcal{D}(G \times G)$ admits partitions of unity [3, pp. 46–49].

3) $\Leftarrow$: Let $\varphi, \psi \in \mathcal{D}(G)$ and $u \in G$ such that $\varphi = 0$ on uP^{-1} and $\operatorname{supp} \psi \subset u(\operatorname{int} P)^{-1}$. Then $\operatorname{supp}(\varphi \otimes \psi) = (\operatorname{supp} \varphi) \times (\operatorname{supp} \psi) \subset Q$, and hence $N\varphi(\psi) = S(\varphi \otimes \psi) = 0$. Q.E.D.

Now we are able to prove the following.

THEOREM 3. *Let G be a locally compact abelian group and let $\Gamma \subset G$ be a closed subgroup such that G/Γ is compact. Assume that the linear system N: $\mathcal{D}(G) \to \mathcal{D}'(G)$ is represented according to Theorem 2 as $N\varphi = \sum_{\chi \in \Gamma^{\perp}} \chi [V_\chi * \varphi]$ for all $\varphi \in \mathcal{D}(G)$. Then N is P-causal if and only if* $\operatorname{supp} V_\chi \subset P$ *for all $\chi \in \Gamma^{\perp}$, that means, if and only if the individual systems N_χ: $N_\chi(\varphi) = V_\chi * \varphi$ for all $\varphi \in \mathcal{D}(G)$ are P-causal for all $\chi \in \Gamma^{\perp}$.*

Proof. 1) Let N be P-causal with corresponding distribution $S \in \mathcal{D}'(G \times G)$. We retain the notations introduced in the proof of Theorem 2. Lemma 3 shows that $\operatorname{supp} S \subset \{(s,t)\colon s,t \in G,\ s^{-1}t \in P\}$ and hence $\operatorname{supp} W \subset P \times G$. It follows that $V_\chi(\varphi) = W(\varphi \otimes \beta\bar{\chi}) = 0$ for all $\varphi \in \mathcal{D}(G)$ with $\operatorname{supp} \varphi \subset G \setminus P$ and for all $\chi \in \Gamma^{\perp}$.

2) It remains to prove the following: If $V \in \mathcal{D}'(G)$ is a distribution with $\operatorname{supp} V \subset P$, then the linear system $\varphi \mapsto V * \varphi$ for all $\varphi \in \mathcal{D}(G)$ is P-causal. To see this, fix $\varphi, \psi \in \mathcal{D}(G)$ and $u \in G$ satisfying $\varphi = 0$ on uP^{-1} and $\operatorname{supp} \psi \subset u(\operatorname{int} P)^{-1}$. We have to show that $\int (V * \varphi) \psi \, dm_G = 0$. Write $x \in \operatorname{supp} \psi$ as $x = uv^{-1}$ with $v \in \operatorname{int} P$. Then $\operatorname{supp}(\tau_x \check{\varphi}) = x(\operatorname{supp} \varphi)^{-1} \subset G \setminus (v^{-1} \operatorname{int} P) \subset G \setminus P$ in view of $vP \subset \operatorname{int} P$. Now the assumption on V tells us that $V * \varphi(x) = V(\tau_x \check{\varphi}) = 0$ for all $x \in \operatorname{supp} \psi$ and hence the assertion. Q.E.D.

Finally let us remark that a regular system N: $\mathcal{D}(G) \to \mathcal{C}(G)$ is P-causal if and only if the function Z: $G \to \mathcal{D}'(G)$ introduced in §4 satisfies $\operatorname{supp} Z(x) \subset P$ for all $x \in G$.

REFERENCES

[1] K. BARBEY, W. HACKENBROCH AND H. WILLIE, *Partially translation invariant linear systems*, Integral Equations and Operator Theory, 3 (1980), pp. 311–322.

[2] R. K. BOSE, *Realizability theory of continuous linear operators on groups*, this Journal, 10 (1979), pp. 767–777.

[3] F. BRUHAT, *Distributions sur un groupe localement compact et applications à l'étude des représentations des groupes p-adiques*, Bull. Soc. Math. France, 89 (1961), pp. 43–75.

[4] P. L. FALB AND M. J. FREEDMAN, *A generalized transform theory for causal operators*, SIAM J. Control, 7 (1969), pp. 452–471.

[5] Y. FOURES AND I. E. SEGAL, *Causality and analyticity*, Trans. Amer. Math. Soc., 78 (1955), pp. 385–405.

[6] A. GROTHENDIECK, *Produits tensoriels topologiques et espaces nucléaires*, Mem. Amer. Math. Soc. 16, American Mathematical Society, Providence, R.I., 1966.

[7] W. HACKENBROCH, *Integraldarstellung einer Klasse dissipativer linearer Operatoren*, Math. Z., 109 (1969), pp. 273–287.

[8] _____, *Passivity and causality over locally compact abelian groups*, unpublished preprint.

[9] G. I. KAC, *Generalized functions on a locally compact group and decompositions of unitary representations*, Trudy Moskov. Mat. Onschch., 10 (1961), pp. 3–40. (In Russian).

[10] H. KÖNIG AND J. MEIXNER, *Lineare Systeme und lineare Transformationen*, Math. Nachr., 19 (1958), pp. 265–322.

[11] K. MAURIN, *Distributionen auf Yamabe-Gruppen. Harmonische Analyse auf einer abelschen l. k. Gruppe*, Bull. Acad. Pol. Sci. Ser. Sci. Math., 9 (1961), pp. 845–850.

[12] D. MONTGOMERY AND L. ZIPPIN, *Topological Transformation Groups*, Interscience, New York, 1966.

[13] A. P. ROBERTSON AND W. J. ROBERTSON, *Topological Vector Spaces*, Cambridge University Press, London, 1966.

[14] R. UNBEHAUEN, *Über das Zeitverhalten linearer Systeme mit sich periodisch ändernden Parametern*, Arch. Elek. Übertr., 23 (1969), pp. 570–574.

[15] R. UNBEHAUEN AND K. FORSTER, *A numerical technique for the time-performance investigation of periodically time-varying systems*, Proc. Biennial Cornell Electrical Engineering Conference on Computerized Electronics, August 1969, pp. 118–129.

[16] H. WILLIE, *Periodisch invariante lineare Übertragungssysteme*, Thesis, Fachbereich Mathematik, Universität Regensburg, West Germany, 1978.

[17] A. H. ZEMANIAN, *Realizability Theory For Continuous Linear Systems*, Academic Press, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0014 $01.25/0

COMPARISON AND STABILITY OF SOLUTIONS FOR A NEUTRON TRANSPORT PROBLEM WITH TEMPERATURE FEEDBACK*

C. V. PAO†

Abstract. A system of coupled equations arising from the neutron transport in a reactor system is investigated where the effect of temperature feedback is taken into consideration. Using the method of successive approximation and the notion of upper and lower solutions, two monotone sequences are constructed for the corresponding integral equations. It is shown that these two sequences converge monotonically from above and below, respectively, to a unique solution of the system. This monotone convergence leads to an existence-comparison theorem in terms of the initial iteration as well as each of the succeeding iterations. Through suitable construction of the initial iteration the existence-comparison theorem is then used to investigate the stability and instability property of a steady-state solution. Sufficient conditions in terms of the physical parameters are given to ensure the stability and instability of the system, including some explicit stability and instability regions. It is also shown under suitable conditions on the same set of physical parameters that global solutions exist for one class of initial functions while they blow up in finite time for another class of initial functions. Characterizations of these two classes of initial functions are obtained. In some special feedback model, global solutions exist for all initial functions but they grow at a rate no less than the order of $\exp(\exp(\eta t))$ for some $\eta>0$.

1. Introduction. In the theory of neutron transport in a nuclear reactor, if the effect of temperature feedback is taken into consideration then the neutron transport equation for the neutron density is supplemented by a temperature equation. Suppose the temperature feedback is only through the multiplication factor in a monoenergetic slab medium where the two faces of the slab are located at $x=0$ and $x=l$. Then according to the balance relation for neutron density and Newton's law of cooling for temperature where the effect of heat conduction is small the equations governing the density function $N=N(t,x,\mu)$ and temperature $T=T(t,x)$ at time t, position x and direction cosine $\mu=\cos\theta$ are given by (cf. [1], [2]):

$$
\begin{aligned}
&v^{-1}N_t+\mu N_x+\sigma_0 N=\left(\frac{\gamma(T)}{2}\right)\int_{-1}^{1}N(t,x,\mu')\,d\mu',\\
&T_t+\beta(T-T_c)=\left(\frac{h(T)}{2l}\right)\int_{-1}^{1}\int_{0}^{l}N(t,x',\mu')\,dx'\,d\mu',\\
&\qquad\qquad (t<0,0<x<l,-1\le\mu\le 1).
\end{aligned}
\tag{1.1}
$$

Here v,σ_0, T_c, β^{-1} are given positive constants representing the neutron speed, total cross-section, coolant temperature and mean time for heat transfer to coolant, respectively, and the functions γ,h, which in general depend on T, are the multiplication factor and the energy generation coefficient. In conventional notations the multiplication factor may be expressed as $\gamma=\sigma_s+\nu\sigma_f$, where σ_s, σ_f are the respective scattering and fission cross-sections and ν is the mean number of secondary neutrons per fission. Assume there is no neutron entering the slab from the outside. Then the boundary condition for the neutron density is given by

$$
N(t,0,\nu)=0 \quad \text{for } t>0,\quad 0<\mu\le 1,\qquad N(t,l,\mu)=0 \quad \text{for } t>0,\quad -1\le\mu<0. \tag{1.2}
$$

As usual, the initial conditions for N,T are in the form

$$
N(0,x,\mu)=N_0(x,\mu),\qquad T(0,x)=T_0(x)\qquad (0<x<l,-1\le\mu\le 1). \tag{1.3}
$$

*Received by the editors December 23, 1980, and in revised form October 3, 1981.
†Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

The system (1.1)–(1.3) gives a mathematical description of the neutron's density and temperature distribution in a slab reactor system. In this system the temperature feedback is considered only through the multiplication factor γ and the energy production coefficient h. In general, the total cross-section σ_0 is also temperature dependent. However, in certain special situations such as the case of prompt feedback due to the Doppler effect on resonance absorption the fission cross-section σ_f is significantly temperature dependent (cf. [1], [8]). For these types of models it is reasonable to consider constant σ_0. On the other hand, the dependence of γ, h on temperature is often determined by the relation

$$\begin{aligned} &\frac{d\gamma}{dT}=a_1T^{-m}, \qquad \gamma(T_c)=\gamma_0>0 \qquad (m\geq 0),\\ &\frac{dh}{dT}=a_2T^{-n}, \qquad h(T_c)=h_0>0 \qquad (n\geq 0) \end{aligned} \tag{1.4}$$

where a_1, a_2 are some constants (cf. [5], [8]). This leads to the explicit expression for γ, h given by

$$\begin{aligned} &\gamma_1(T)=\gamma_0+a_1\ln\left(\frac{T}{T_c}\right) \qquad (m=1),\\ &\gamma_m(T)=\gamma_0-a_1(m-1)\left(T^{-(m-1)}-T_c^{-(m-1)}\right) \qquad (m\neq 1),\\ &h_1(T)=h_0+a_2\ln\left(\frac{T}{T_c}\right) \qquad (n=1),\\ &h_n(T)=h_0-a_2(n-1)\left(T^{-(n-1)}-T_c^{-(n-1)}\right) \qquad (n\neq 1). \end{aligned} \tag{1.5}$$

Of special interest is the function $\gamma_m(T)$ in (1.5) with $m=3/2$ (see [8]). In this paper we only consider the case of positive temperature feedback $a_i\geq 0$ $(i=1,2)$ which is an important concern in reactor stability consideration.

The neutron transport equation has been investigated by many researchers in the field, but most of the discussions are devoted to the linear transport equation where the effect of temperature feedback is neglected (e.g., see [4], [9], [14]). When the temperature effect is taken into consideration it is often investigated in the framework of diffusion approximation, especially in relation to the qualitative property of the solution (cf. [6]–[8], [11]–[13]). An important problem in these investigations is to determine how the temperature affects the stability property of the flux distribution in the reactor system. The work in [11] also discusses the blowing-up property of the solution. On the other hand, the neutron transport problem with temperature feedback without the diffusion approximation has been investigated in [2], [3]. The main concern in these papers is the existence-uniqueness question using semi-group theory of evolution equations. As usual, the definition of stability and instability in this paper is in the sense of Lyapunov. For physical reasons, our stability or instability of the steady-state solution $(0, T_c)$ is always with respect to nonnegative initial perturbations $(N_0, T_0)\geq(0, T_c)$.

The purpose of this paper is twofold: first we develop an iterative scheme for the construction of a solution from which an existence-comparison theorem is established. An important consequence of this theorem is that it gives an upper and a lower bound of the solution in terms of the initial iteration. Furthermore, each succeeding iteration narrows the gap between the upper and the lower bound of the solution. Our second goal is to use the existence-comparison theorem to investigate the asymptotic behavior of the solution and the possible blowing-up property of the system. Sufficient conditions in terms of the physical parameters are given to ensure the stability, instability

and blowing-up behavior of the system without explicit knowledge of the solution. In the case of a given finite time interval, both upper and lower bounds of the solution are obtained.

The plan of the paper is as follows: in §2, we construct two monotone sequences and show that these sequences converge monotonically from above and below, respectively, to a unique solution of the system. This monotone convergence leads to an existence-comparison theorem in terms of the initial iteration as well as succeeding iterations. A recursion formula for the calculation of the iterations is included in the discussion. Through suitable construction of the initial iterations, called upper and lower solutions, we investigate in §3 the stability and instability property of the steady-state solution $(0, T_c)$. Sufficient conditions for the stability and instability as well as a stability and an instability region of $(0, T_c)$ are explicitly given. Finally, in §4 we discuss the existence and nonexistence of global solutions. Special attention is given to the functions γ, h given by (1.5). It is shown for this model that if $m \geq 1$, $n \geq 1$, then for every nonnegative initial function (N_0, u_0), a unique global solution always exists. Furthermore, the solution grows no faster than an exponential order when $m > 1$, $n > 1$ but it may grow in the order of $\exp(\exp(\eta t))$ for some $\eta > 0$ when $m = n = 1$. However, if $m < 1$, $n < 1$ then global solutions exist for one class of initial functions while they blow up in finite time for another class of initial functions. Estimates for these two classes of initial functions are obtained.

2. The existence-comparison theorem. Let $D_1 = (0, t_1] \times (0, l)$, $D_2 = (0, l) \times [-1, 1]$, $Q = (0, t_1] \times (0, l) \times [-1, 1]$ and let $\bar{D}_i$, $\bar{Q}$ denote the respective closure of D_i and Q, where $t_1 > 0$ is finite but can be arbitrarily large. Throughout this paper we assume that γ, h are positive continuous functions on $R^+ \equiv [0, \infty)$. The aim of this section is to establish an existence-comparison theorem by constructing two monotone sequences which converge from above and below, respectively, to a unique solution of (1.1)–(1.3). For this purpose, it is convenient to let $u = T - T_c$ and transform the system (1.1)–(1.3) into the form

$$
\begin{aligned}
&v^{-1} N_t + \mu N_x + \sigma_0 N = \frac{\gamma(u)}{2} \int_{-1}^{1} N(t, x, \mu')\, d\mu' \qquad ((t, x, \mu) \in Q), \\
&u_t + \beta u = \frac{h(u)}{2l} \int_{-1}^{1} \int_{0}^{l} N(t, x', \mu')\, dx'\, d\mu' \qquad ((t, x) \in D_1),
\end{aligned} \tag{2.1}
$$

$$
N(t, 0, \mu) = 0 \quad (t < 0, 0 < \mu \leq 1), \qquad N(t, l, \mu) = 0 \quad (t > 0, -1 \leq \mu < 0), \tag{2.2}
$$

$$
N(0, x, \mu) = N_0(x, \mu), \quad u(0, x) = u_0(x) \qquad ((x, \mu) \in D_2), \tag{2.3}
$$

where $u_0(x) = T_0(x) - T_c(x)$. By considering $(N_t + v\mu N_x)$ as the total derivative $(d/dt) N(t, x + v\mu t)$, an integration of (2.1), using the conditions (2.2), (2.3), yields the integral equation

(2.4)

$$
\begin{aligned}
N(t, x, \mu) &= \exp(-v\sigma_0 t) N_0(x - v\mu t, \mu) \\
&\quad + \int_0^t \exp[-v\sigma_0 (t - \tau)] (f_1(N, u))(\tau, x - v\mu(t - \tau))\, d\tau \qquad ((t, x, \mu) \in \bar{Q}),
\end{aligned}
$$

$$
\begin{aligned}
u(t, x) &= u_0(x) \exp(-\beta t) + \int_0^t \exp[-\beta(t - \tau)] (f_2(N, u))(\tau, x)\, d\tau \\
&\qquad ((t, x) \in \bar{D}_1),
\end{aligned} \tag{2.5}
$$

where

$$(f_1(N,u))(t,x)=\frac{\gamma(u(t,x))}{2}\int_{-1}^{1}N(t,x,\mu')\,d\mu',$$
$$(f_2(N,u))(t,x)=\frac{h(u(t,x))}{2l}\int_{-1}^{1}\int_{0}^{l}N(t,x',\mu')\,dx'\,d\mu'. \tag{2.6}$$

In the integral representation (2.4) it is defined that

$$N_0(x,\mu)=(f_1(N,u))(t,x)=0 \quad \text{when } x\notin[0,l].$$

Clearly, every solution of the differential system (2.1)–(2.3) is a solution of the integral equations (2.4), (2.5). Conversely, every solution of (2.4), (2.5) is also a solution of (2.1)–(2.3) when $(N_t+v\mu N_x)$ is considered as the total derivative $(d/dt)N(t,x+v\mu t)$. In this paper, a solution of (2.1)–(2.3) (or (1.1)–(1.3)) is always meant in the above sense. It is to be noted that in the present definition of a solution the function (N,u) needs to be continuous and possesses a total derivative in t (in the classical sense). Hence the solution defined in this paper is stronger than the solution in the L^p-space for every p $(1\le p\le\infty)$. For physical reasons, we are concerned with only nonnegative solutions of (2.1)–(2.3).

To ensure the existence of a global nonnegative solution of (2.1)–(2.3) it is necessary to impose some conditions on γ,h. Our basic hypothesis on these functions is the following.

(H_0). There exist positive constants ρ,K such that

$$\gamma(u_2)\ge\gamma(u_1)>0$$
$$|h(u_2)-h(u_1)|\le K|u_2-u_1| \qquad (0\le u_1\le u_2\le\rho). \tag{2.7}$$

The requirement on γ in (H_0) corresponds to positive temperature feedback which is an important concern in reactor stability consideration. In the special case where γ is given by (1.5) this requirement is fulfilled for every $\rho<\infty$ when $a_1\ge 0$. In general, condition (2.7) is required only for some $\rho>0$, and the value of ρ is often determined by the magnitude of upper and lower solutions which are defined as follows:

DEFINITION 2.1. A continuous vector-valued function $(\tilde N,\tilde u)$ is called an *upper solution* of (2.4), (2.5) if it satisfies

$$\tilde N(t,x,\mu)\ge\exp(-v\sigma_0 t)N_0(x-v\mu t,\mu)+\int_0^t\exp[-v\sigma_0(t-\tau)]\big(f_1(\tilde N,\tilde u)\big)(\tau,x-v(t-\tau))\,d\tau,$$
$$\tilde u(t,x)\ge u_0(x)\exp(-\beta t)+\int_0^t\exp[-\beta(t-\tau)]\big(f_2(\tilde N,\tilde u)\big)(\tau,x)\,d\tau. \tag{2.8}$$

Similarly, a continuous function $(\underset{\sim}{N},\underset{\sim}{u})$ is called a *lower solution* if it satisfies the reversed inequalities in (2.8).

Let $(\tilde N,\tilde u)$, $(\underset{\sim}{N},\underset{\sim}{u})$ be upper and lower solutions with $(\tilde N,\tilde u)\ge(\underset{\sim}{N},\underset{\sim}{u})\ge(0,0)$ (i.e., $\tilde N\ge\underset{\sim}{N}\ge 0,\tilde u\ge\underset{\sim}{u}\ge 0$) and let ρ,M be positive constants such that $\rho\ge\tilde u$, $M\ge K\tilde N$ on $\overline{Q}$, where ρ,K are the constants in (H_0). Then by adding Mu on both sides of the second equation in (2.1) and integrating we obtain the corresponding integral equation

$$u(t,x)=u_0(x)\exp(-(\beta+M)t)+\int_0^t\exp[-(\beta+M)(t-\tau)]\big(Mu+f_2(N,u)\big)(\tau,x)\,d\tau \qquad \big((t,x)\in\overline{D}_1\big). \tag{2.9}$$

It is clear that this integral equation coincides with the equation in (2.5). By using $(N^{(0)},u^{(0)})=(\tilde{N},\tilde{u})$,as an initial iteration we can construct a sequence, denoted by $\{\bar{N}^{(k)},\bar{u}^{(k)}\}$, from the recursion formula

(2.10)

$$N^{(k)}(t,x,\mu)=\exp(-v\sigma_0 t)N_0(x-v\mu t,\mu)+\int_0^t \exp[-v\sigma_0(t-\tau)]\big(f_1(N^{(k-1)},u^{(k-1)})\big)(\tau,x-v\mu(t-\tau))\,d\tau,$$

$$u^{(k)}(t,x)=u_0(x)\exp[-(\beta+M)t]+\int_0^t \exp[-(\beta+M)(t-\tau)]\big(Mu^{(k-1)}+f_2(N^{(k-1)},u^{(k-1)})\big)(\tau,x)\,d\tau \qquad (k=1,2,\cdots).$$

Similarly, we can obtain another sequence from (2.10) by using the initial iteration $(N^{(0)},u^{(0)})=(\underset{\sim}{N},\underset{\sim}{u})$, and this is denoted by $\{\underline{N}^{(k)},\underline{u}^{(k)}\}$. These two sequences, referred to as maximal and minimal sequence, respectively, possess the following monotone properties.

LEMMA 2.1. *The maximal sequence* $\{\bar{N}^{(k)},\bar{u}^{(k)}\}$ *is monotone nonincreasing and the minimal sequence* $\{\underline{N}^{(k)},\underline{u}^{(k)}\}$ *is monotone nondecreasing. Moreover,*

$$(\underline{N}^{(k)},\underline{u}^{(k)})\le(\bar{N}^{(k)},\bar{u}^{(k)}) \quad \text{on } \bar{Q} \text{ for every } k=1,2,\cdots.$$

Proof. It is easily seen by letting $k=1$ in (2.10) and using the definition of upper and lower solutions that $(\bar{N}^{(1)},\bar{u}^{(1)})\le(\bar{N}^{(0)},\bar{u}^{(0)})$, $(\underline{N}^{(1)},\underline{u}^{(1)})\ge(\underline{N}^{(0)},\underline{u}^{(0)})$. We show that $(\bar{N}^{(1)},\bar{u}^{(1)})\ge(\underline{N}^{(1)},\underline{u}^{(1)})$. Let (N_1,u_1), (N_2,u_2) be any two functions such that $(0,0)\le(N_1,u_1)\le(N_2,u_2)\le(\tilde{N},\tilde{u})$. Then by (2.6), (2.7) and the nonnegative property of $h(u)$,

$$f_1(N_2,u_2)-f_1(N_1,u_1)=\frac{\gamma(u_2)}{2}\int_{-1}^1 N_2\,d\mu'-\frac{\gamma(u_1)}{2}\int_{-1}^1 N_1\,d\mu'\ge 0,$$

$$f_2(N_2,u_2)-f_2(N_1,u_1)=\frac{h(u_2)}{2l}\int_{-1}^1\int_0^l N_2\,dx'\,d\mu'-\frac{h(u_1)}{2l}\int_{-1}^1\int_0^l N_1\,dx'\,d\mu' \ge\left[\frac{-K}{2l}\int_{-1}^1\int_0^l N_2\,dx'\,d\mu'\right](u_2-u_1), \tag{2.11}$$

where we have suppressed the dummy variables in the integral expressions. Since $M\ge K\tilde{N}\ge(K/2l)\int_{-1}^1\int_0^l N_2\,dx'\,d\mu'$ and $(0,0)\le(\underline{N}^{(0)},\underline{u}^{(0)})\le(\bar{N}^{(0)},\bar{u}^{(0)})=(\tilde{N},\tilde{u})$ the above relation and (2.10) imply that

$$\bar{N}^{(1)}-\underline{N}^{(1)}=\int_0^t \exp[-v\sigma_0(t-\tau)]\big(f_1(\bar{N}^{(0)},\bar{u}^{(0)})-f_1(\underline{N}^{(0)},\underline{u}^{(0)})\big)\cdot(\tau,x-v\mu(t-\tau))\,d\tau\ge 0,$$

$$\bar{u}^{(1)}-\underline{u}^{(1)}=\int_0^t \exp[(\beta+M)(t-\tau)]\big[M(\bar{u}^{(0)}-\underline{u}^{(0)})+f_2(\bar{N}^{(0)},\bar{u}^{(0)})-f_2(\underline{N}^{(0)},\underline{u}^{(0)})\big](\tau,x)\,d\tau\ge 0. \tag{2.12}$$

This proves the relation

$$(\underline{N}^{(0)},\underline{u}^{(0)})\le(\underline{N}^{(1)},\underline{u}^{(1)})\le(\bar{N}^{(1)},\bar{u}^{(1)})\le(\bar{N}^{(0)},\bar{u}^{(0)}).$$

Assume, by induction, that

$$(\underline{N}^{(k-1)},\underline{u}^{(k-1)})\le(\underline{N}^{(k)},\underline{u}^{(k)})\le(\overline{N}^{(k)},\overline{u}^{(k)})\le(\overline{N}^{(k-1)},\overline{u}^{(k-1)}). \tag{2.13}$$

Then by (2.10),(2.11) with $(N_2,u_2)=(\overline{N}^{(k-1)},\overline{u}^{(k-1)})$, $(N_1,u_1)=(\overline{N}^{(k)},\overline{u}^{(k)})$ we have

$$\overline{N}^{(k)}-\overline{N}^{(k+1)}=\int_0^t \exp[-v\sigma_0(t-\tau)]\big(f_1(\overline{N}^{(k-1)},\overline{u}^{(k-1)})-f_1(\overline{N}^{(k)},\overline{u}^{(k)})\big)$$
$$\cdot(\tau,x-v\mu(t-\tau))\,d\tau\ge 0,$$

$$\overline{u}^{(k)}-\overline{u}^{(k+1)}=\int_0^t \exp[-(\beta+M)(t-\tau)]$$
$$\cdot\big[M(\overline{u}^{(k-1)}-\overline{u}^{(k)})+f_2(\overline{N}^{(k-1)},\overline{u}^{(k-1)})-f_2(\overline{N}^{(k)},\overline{u}^{(k)})\big](x,\tau)\,d\tau\ge 0$$

which proves the relation $(\overline{N}^{(k)},\overline{u}^{(k)})\ge(\overline{N}^{(k+1)},\overline{u}^{(k+1)})$. A similar argument leads to the conclusion $(\underline{N}^{(k)},\underline{u}^{(k)})\le(\underline{N}^{(k+1)},\underline{u}^{(k+1)})$ and $(\underline{N}^{(k)},\underline{u}^{(k)})\le(\overline{N}^{(k)},\overline{u}^{(k)})$. It follows from the inductive principle that (2.13) holds for every k. This proves the conclusion of the lemma.

In view of Lemma 2.1 the pointwise (and componentwise) limits

$$\lim_{k\to\infty}(\overline{N}^{(k)},\overline{u}^{(k)})=(\overline{N},\overline{u}),\qquad \lim_{k\to\infty}(\underline{N}^{(k)},\underline{u}^{(k)})=(\underline{N},\underline{u}) \tag{2.14}$$

exist and the convergence of these sequences are monotone. Letting $k\to\infty$ in (2.10) and applying the dominated convergence theorem we see that both $(\overline{N},\overline{u})$ and $(\underline{N},\underline{u})$ are solutions of the integral equations (2.4), (2.9). The equivalence between the equations in (2.5) and (2.9) ensures that they are also solutions of (2.4), (2.5). This conclusion leads to the following existence-comparison theorem.

THEOREM 2.1. *Let* $(\tilde{N},\tilde{u})$, $(\underset{\sim}{N},\underset{\sim}{u})$ *be upper and lower solutions such that* $(\tilde{N},\tilde{u})\ge(\underset{\sim}{N},\underset{\sim}{u})\ge(0,0)$ *and let* (H_0) *hold for some* $\rho\ge\tilde{u}$. *Then the maximal sequence* $\{\overline{N}^{(k)},\overline{u}^{(k)}\}$ *converges montonically to a solution* $(\overline{N},\overline{u})$ *of* (2.1)–(2.3) *and the minimal sequence* $\{\underline{N}^{(k)},\underline{u}^{(k)}\}$ *converges monotonically to a solution* $(\underline{N},\underline{u})$. *Moreover,*

$$(\underset{\sim}{N},\underset{\sim}{u})\le(\underline{N}^{(k)},\underline{u}^{(k)})\le(\underline{N},\underline{u})\le(\overline{N},\overline{u})\le(\overline{N}^{(k)},\overline{u}^{(k)})\le(\tilde{N},\tilde{u}) \tag{2.15}$$

for every $k=1,2,\cdots$.

In order to show that $(\overline{N},\overline{u})$ coincides with $(\underline{N},\underline{u})$ and is the unique solution of (2.1)–(2.3) we need to impose a Lipschitz condition on γ.

(H_1). There exist positive constants K_1, ρ such that

$$|\gamma(u_2)-\gamma(u_1)|\le K_1|u_2-u_1| \quad \text{for } 0\le u_1\le u_2\le\rho. \tag{2.16}$$

Again the Lipschitz condition on γ is required only on the finite interval $[0,\rho]$. With this additional condition we have the following uniqueness-comparison theorem.

THEOREM 2.2. *Let* $(\tilde{N},\tilde{u})$, $(\underset{\sim}{N},\underset{\sim}{u})$ *be upper and lower solutions such that* $(\tilde{N},\tilde{u})\ge(\underset{\sim}{N},\underset{\sim}{u})\ge(0,0)$ *and let* (H_0), (H_1) *hold for some* $\rho\ge\tilde{u}$. *Then* $(\overline{N},\overline{u})=(\underline{N},\underline{u})$ *and is the unique solution* (N,u) *of* (2.1)–(2.3) *such that*

$$(\underset{\sim}{N},\underset{\sim}{u})\le(\underline{N}^{(k)},\underline{u}^{(k)})\le(N,u)\le(\overline{N}^{(k)},\overline{u}^{(k)})\le(\tilde{N},\tilde{u}) \tag{2.17}$$

for every $k=1,2,\cdots$.

Proof. Let λ be an arbitrary positive constant to be chosen and let

$$N^*=e^{-\lambda t}(\overline{N}-\underline{N}),\qquad u^*=e^{-\lambda t}(\overline{u}-\underline{u}).$$

Then $(N^*,u^*)\geq(0,0)$ and satisfies the equations

$$(N_t^*+v\mu N_x^*)+(v\sigma_0+\lambda)N^*=\frac{ve^{-\lambda t}}{2}\left[\gamma(\bar{u})\int_{-1}^{1}\bar{N}\,d\mu'-\gamma(\underline{u})\int_{-1}^{1}\underline{N}\,d\mu'\right],$$

$$u_t^*+(\beta+\lambda)u^*=\frac{e^{-\lambda t}}{2l}\left[h(\bar{u})\int_{-1}^{1}\int_{0}^{l}\bar{N}\,dx'\,d\mu'-h(\underline{u})\int_{-1}^{1}\int_{0}^{l}\underline{N}\,dx'\,d\mu'\right], \tag{2.18}$$

$$N^*(t,0,\mu)=0\quad(t>0,0<\mu\leq 1),\qquad N^*(t,l,\mu)=0\quad(t>0,-1\leq\mu<0),$$

$$N^*(0,x,\mu)=0,\qquad u^*(0,x)=0\qquad((x,\mu)\in D_2).$$

Denote by $\bar{h}$ the least upper bound of $h(u)$ for $0\leq u\leq\rho$ and let M_1, M_2 be any positive constants satisfying $2v^{-1}M_1\geq\gamma(\tilde{u})+K_1\int_{-1}^{1}\tilde{N}\,d\mu'$, $2lM_2\geq\bar{h}+K\int_{-1}^{1}\int_{0}^{l}\tilde{N}\,dx'\,d\mu'$, where K_1, K are the Lipschitz constants of γ,h. Then by the hypotheses (H_0), (H_1),

$$\begin{aligned}\gamma(\bar{u})\int_{-1}^{1}\bar{N}\,d\mu'-\gamma(\underline{u})\int_{-1}^{1}\underline{N}\,d\mu'&\leq\left(K_1\int_{-1}^{1}\bar{N}\,d\mu'\right)(\bar{u}-\underline{u})+\gamma(\underline{u})\int_{-1}^{1}(\bar{N}-\underline{N})\,d\mu'\\&\leq 2v^{-1}M_1e^{\lambda t}\left(u^*+\int_{-1}^{1}N^*\,d\mu'\right),\\h(\bar{u})\int_{-1}^{1}\int_{0}^{l}\bar{N}\,dx'\,d\mu'-h(\underline{u})\int_{-1}^{1}\int_{0}^{l}\underline{N}\,dx'\,d\mu'&\leq\left(K\int_{-1}^{1}\int_{0}^{l}\bar{N}\,dx'\,d\mu'\right)(\bar{u}-\underline{u})\\&\quad+h(\underline{u})\int_{-1}^{1}\int_{0}^{l}(\bar{N}-\underline{N})\,dx'\,d\mu'\\&\leq 2lM_2e^{\lambda t}\left(u^*+\int_{-1}^{1}\int_{0}^{l}N^*\,dx'\,d\mu'\right).\end{aligned}$$

Using the above relation in the first two equations of (2.18) and letting $\lambda_1=v\sigma_0+\lambda$, $\lambda_2=\beta+\lambda$ we obtain

$$\begin{aligned}&(N_t^*+v\mu N^*{}_x)+\lambda_1N^*\leq M_1\left(u^*+\int_{-1}^{1}N^*\,d\mu'\right),\\&u_t^*+\lambda_2u^*\leq M_2\left(u^*+\int_{-1}^{1}\int_{0}^{l}N^*\,dx'\,d\mu'\right).\end{aligned} \tag{2.19}$$

The above inequalities together with the boundary and initial conditions in (2.18) imply that

$$\begin{aligned}&N^*(t,x,\mu)\leq M_1\int_0^t\left[\exp(-\lambda_1(t-\tau))\left(u^*+\int_{-1}^{1}N^*\,d\mu'\right)(\tau,x-v\mu(t-\tau))\right]d\tau,\\&u^*(t,x)\leq M_2\int_0^t\left[\exp(-\lambda_2(t-\tau))\left(u^*+\int_{-1}^{1}\int_{0}^{l}N^*\,dx'\,d\mu'\right)(\tau,x)\right]d\tau.\end{aligned} \tag{2.20}$$

Define

$$\|N\|=\sup\{|N(t,x,\mu)|;(t,x,\mu)\in Q\},\qquad \|u\|=\sup\{|u(t,x)|;(t,x)\in D_1\}.$$

Then by (2.20) and the nonnegative property of (N^*,u^*),

$$\|N^*\|\leq M_1(\|u^*\|+2\|N^*\|)\int_0^t\exp(-\lambda_1(t-\tau))\,d\tau\leq\frac{M_1}{\lambda_1}(\|u^*\|+2\|N^*\|),$$

$$\|u^*\|\leq M_2(\|u^*\|+2l\|N^*\|)\int_0^t\exp(-\lambda_2(t-\tau))\,d\tau\leq\frac{M_2}{\lambda_2}(\|u^*\|+2l\|N^*\|).$$

Addition of the above two inequalities leads to the relation

$$\|N^*\|+\|u^*\|\le\left(2M_1\lambda_1^{-1}+2lM_2\lambda_2^{-1}\right)\left(\|N^*\|+\|u^*\|\right). \tag{2.21}$$

However, if λ is chosen so that $\lambda_1>4M$, $\lambda_2\ge 4lM_2$, then the relation (2.21) cannot hold unless $\|N^*\|+\|u^*\|=0$. This implies that $\bar{N}=\underline{N}$ and $\bar{u}=\underline{u}$. Now if (N,u) is any other solution such that $(\underline{N},\underline{u})\le(N,u)\le(\tilde{N},\tilde{u})$ then it is also an upper solution as well as a lower solution. Using (N,u) as the initial iteration in (2.10), the same argument as in the proof of Lemma 2.1 shows that $(N,u)\le(\bar{N}^{(k)},\bar{u}^{(k)})$ and $(N,u)\ge(\underline{N}^{(k)},\underline{u}^{(k)})$ for every $k=1,2,\cdots$. It follows that $(\bar{N},\bar{u})\ge(N,u)\ge(\underline{N},\underline{u})$ and therefore $(\bar{N},\bar{u})$ (or $(\underline{N},\underline{u})$) is the unique solution. This completes the proof of the theorem.

Remark 2.1. (a) The conditions in (H_0), (H_1) are all satisfied if γ,h are continuously differentiable in u and γ is nondecreasing in u for $0\le u\le\rho$. The constant ρ is determined by the least upper bound of $\tilde{u}$ and plays an important role in the determination of a stability region.

(b) The uniqueness of the solution in Theorem 2.2 is insured only in the region bounded by upper and lower solutions, and nothing can be said about the system outside this region. However, if the Lipschitz conditions (2.7), (2.16) for h,γ hold for every finite ρ, where the Lipschitz constants K,K_1 may depend on ρ, then there exists exactly one nonnegative solution. This can be seen from the proof of Theorem 2.2. Notice that the uniqueness proof depends on the fact that $(\bar{N},\bar{u})\ge(\underline{N},\underline{u})$.

3. Stability and instability problem. It is seen from Theorem 2.2 that the existence of upper and lower solutions gives not only the existence and upper and lower bounds of the solution, but also that a suitable construction of these functions can often determine the stability and instability property of a steady-state solution. The aim of this section is to construct explicit upper and lower solutions so that either a unique global solution exists and converges to a steady-state solution or it grows unbounded as $t\to\infty$. This decay or growth property of the solution depends on the physical parameters of the system without explicit knowledge of the solution. We first establish the global existence problem when γ,h are uniformly bounded in R^+.

THEOREM 3.1. *Let γ,h satisfy (H_0), (H_1) for every finite $\rho>0$ and let $\gamma(u)\le b_1$, $h(u)\le b_2$ in R^+ for some positive constants b_1, b_2. Then the problem* (2.1)–(2.3) *has a unique global solution (N,u) such that*

$$0\le N(t,x,\mu)\le\rho_1 e^{\alpha t},\qquad 0\le u(t,x)\le\rho_2 e^{\alpha t}\qquad \left(t>0,(x,\mu)\in\bar{D}_2\right) \tag{3.1}$$

whenever $(0,0)\le(N_0,u_0)\le(\rho_1,\rho_2)$, where α,ρ_1,ρ_2 are constants with α (not necessarily positive) satisfying

$$\alpha\ge\max\left\{v(b_1-\sigma_0),\frac{b_2\rho_1}{\rho_2}-\beta\right\}. \tag{3.2}$$

Proof. In view of Theorem 2.2 it suffices to show that $(\tilde{N},\tilde{u})=(\rho_1e^{\alpha t},\rho_2e^{\alpha t})$, $(\underline{N},\underline{u})=(0,0)$ are upper and lower solutions of (2.1)–(2.3), respectively. Indeed, since by (2.6), $f_i(0,0)=0$ for each $i=1,2$, the zero function $(0,0)$ is a lower solution. To show that $(\rho_1e^{\alpha t},\rho_2e^{\alpha t})$ is an upper solution we observe from the equivalence between the differential system (2.1)–(2.3) and the integral representation (2.4), (2.5) that $(\tilde{N},\tilde{u})$ is an upper solution if

$$\begin{aligned}
&v^{-1}\tilde{N}_t+\mu\tilde{N}_x+\sigma_0\tilde{N}\ge f_1(\tilde{N},\tilde{u}),\qquad \tilde{u}_t+\beta\tilde{u}\ge f_2(\tilde{N},\tilde{u}),\\
&\tilde{N}(t,0,\mu)\ge 0\quad(t>0,0<\mu\le 1),\qquad \tilde{N}(t,l,\mu)\ge 0\quad(t>0,-1\le\mu<0),\\
&\tilde{N}(0,x,\mu)\ge N_0(x,\mu),\tilde{u}(0,x)\ge u_0(x).
\end{aligned} \tag{3.3}$$

Since the boundary and initial requirements are clearly satisfied by $(\tilde{N},\tilde{u})=(\rho_1 e^{\alpha t},\rho_2 e^{\alpha t})$ we only need to verify that

$$(v^{-1}\alpha+\sigma_0)\rho_1 e^{\alpha t}\geq\gamma(\rho_2 e^{\alpha t})\rho_1 e^{\alpha t},$$
$$(\alpha+\beta)\rho_2 e^{\alpha t}\geq h(\rho_2 e^{\alpha t})\rho_1 e^{\alpha t}.$$

By the hypothesis $\gamma(u)\leq b_1$, $h(u)\leq b_2$ for all $u\geq 0$, the above inequalities are satisfied by any constant α such that $v^{-1}\alpha+\sigma_0\geq b_1$ and $\alpha+\beta\geq b_2\rho_1/\rho_2$. This leads to the choice of α satisfying (3.2). The existence of a global solution and the relation (3.1) follow from Theorem 2.2.

The result of Theorem 3.1 states that if γ,h are uniformly bounded then for any nonnegative initial function a unique global solution to (2.1)–(2.3) exists and grows no faster than an exponential order. In fact, if $b_1<\sigma_0$, $b_2\rho_1/\rho_2<\beta$ then α can be taken negative and the solution decays exponentially to $(0,0)$ as $t\to\infty$. Since the condition $b_2\rho_1/\rho_2<\beta$ can always be satisfied by taking ρ_2 sufficiently large the decayed property of the solution is insured whenever $b_1<\sigma_0$. This is, of course, to be expected physically. On the other hand, if one or both of the functions γ,h are not uniformly bounded, global solution may not exist, and even if it exists it may not converge to a steady-state solution. In this situation it is important to know under what condition on γ,h and for what class of initial functions the corresponding global solutions exist and converge to a steady state. It is also interesting to know when the solutions grow unbounded and at what rate. In order to investigate these questions in the framework of the previous section it is necessary to construct some different upper and lower solutions. To this end we first find a function $\phi(x,\mu)\geq 0$ satisfying the equations

$$\begin{aligned}&\mu\phi_x+\sigma_0\phi=M_0 \qquad ((x,\mu)\in D_2),\\ &\phi(0,\mu)=0 \quad \text{for } 0<\mu\leq 1, \qquad \phi(l,\mu)=0 \quad \text{for } -1\leq\mu<0,\end{aligned} \tag{3.4}$$

where M_0 is a given positive constant. It is easily seen that the solution of the above boundary-value problem is given by (cf. [9], [10])

$$\phi(x,\mu)=\begin{cases}\dfrac{M_0}{\sigma_0}(1-e^{-\sigma_0 x/\mu}) & (\mu>0),\\ \dfrac{M_0}{\sigma_0} & (\mu=0),\\ \dfrac{M_0}{\sigma_0}(1-e^{\sigma_0(l-x)/\mu}) & (\mu<0).\end{cases} \tag{3.5}$$

We choose the constant M_0 so that ϕ is normalized in the sense that $\int_{-1}^{1}\int_0^l\phi(x,\mu)\,dx\,d\mu=2l$. Direct integration of ϕ shows that M_0 is given by

$$M_0=\frac{\sigma_0}{2l}\left[\int_0^l\int_0^1(2-e^{-\sigma_0 x/\mu}-e^{-\sigma_0(l-x)/\mu})\,d\mu\,dx\right]^{-1}. \tag{3.6}$$

We next define

$$\begin{aligned}&\bar{h}(u_0)=\sup\{h(u);\,0\leq u\leq u_0\}, \qquad \underline{h}(u_0)=\inf\{h(u);\,u\geq u_0\},\\ &E_2(z)=\int_0^1\exp(-z/\mu)\,d\mu \qquad (z\geq 0),\end{aligned} \tag{3.7}$$

where for simplicity, u_0 is taken as a constant. The function $E_2(z)$ is the exponential integral of order two. With these notations we have the following existence-stability result when γ,h are not necessarily bounded.

THEOREM 3.2. *Let γ, h satisfy* (H_0), (H_1) *for some $\rho \geq u_0 \geq 0$, where u_0 is a constant. If*

$$\frac{\sigma_0}{\gamma(u_0)} > 1 - E_2\left(\frac{\sigma_0 l}{2}\right) \tag{3.8}$$

then for any $N_0 \leq \rho_0 \phi$ with $\rho_0 < \beta u_0 / \bar{h}(u_0)$ there exists a constant $\alpha > 0$ such that a unique global solution (N, u) exists and satisfies

$$\begin{aligned} &N_0(x - v\mu t, \mu) e^{-\alpha t} \leq N(t, x, \mu) \leq \rho_0 e^{-\alpha t} \phi(x, \mu) \\ &u_0 e^{-(\beta + M)t} \leq u(t, x) \leq u_0 e^{-\alpha t} \end{aligned} \qquad (t > 0,\ (x, \mu) \in \bar{D}_2) \tag{3.9}$$

where $M = K\rho_0$ and K is the Lipschitz constant in (2.7).

Proof. We first show that $(\tilde{N}, \tilde{u}) = (\rho_0 e^{-\alpha t}\phi, u_0 e^{-\alpha t})$, $(\underline{N}, \underline{u}) = (0, 0)$ are upper and lower solutions, respectively. It is clear from the proof of Theorem 3.1 that $(0, 0)$ is a lower solution. Since $\tilde{N}(0, x, \mu) = \rho_0 \phi \geq N_0$, $\tilde{u}(0, x) = u_0$, and by (3.4), $\tilde{N}(t, 0, \mu) = 0$ for $0 < \mu \leq 1$, $\tilde{N}(t, l, \mu) = 0$ for $-1 \leq \mu < 0$, we only need to show that $(\tilde{N}, \tilde{u})$ satisfies the differential inequalities in (3.3), that is,

$$\left[\mu \phi_x + (\sigma_0 - v^{-1}\alpha)\phi\right] \rho_0 e^{-\alpha t} \geq \frac{\gamma(u_0 e^{-\alpha t})}{2} \int_{-1}^{1} \rho_0 e^{-\alpha t} \phi(x, \mu')\, d\mu',$$

$$(\beta - \alpha) u_0 e^{-\alpha t} \geq \frac{h(u_0 e^{-\alpha t})}{2l} \int_{-1}^{1} \int_{0}^{l} \rho_0 e^{-\alpha t} \phi(x', \mu')\, dx'\, d\mu'.$$

In view of the equation (3.4) and the normalized property of ϕ the above relation is equivalent to

$$M_0 - v^{-1}\alpha\phi \geq \frac{\gamma(u_0 e^{-\alpha t})}{2} \int_{-1}^{1} \phi(x, \mu')\, d\mu',$$

$$(\beta - \alpha) u_0 \geq \rho_0 h(u_0 e^{-\alpha t}).$$

It is clear from $\rho_0 < \beta u_0 / \bar{h}(u_0)$ that the second inequality is satisfied by a sufficiently small $\alpha > 0$. Since $\gamma(u_0 e^{-\alpha t}) \leq \gamma(u_0)$ and ϕ is given by (3.5) the first inequality is also satisfied when

$$\begin{aligned} M_0 - v^{-1}\alpha\phi &\geq \frac{M_0 \gamma(u_0)}{2\sigma_0} \left[\int_0^1 (1 - e^{-\sigma_0 x/\mu'})\, d\mu' + \int_{-1}^{0} (1 - e^{\sigma_0 (l-x)/\mu'})\, d\mu'\right] \\ &= \frac{M_0 \gamma(u_0)}{2\sigma_0} \omega(x), \end{aligned} \tag{3.10}$$

where

$$\omega(x) = \int_0^1 [2 - e^{-\sigma_0 x/\mu'} - e^{-\sigma_0 (l-x)/\mu'}]\, d\mu'. \tag{3.11}$$

It is easily seen that the maximum value of $\omega(x)$ occurs at $x = l/2$ and is given by $\omega(l/2) = 2(1 - E_2(\sigma_0 l/2))$. Hence condition (3.10) holds if

$$M_0 - v^{-1}\alpha\phi \geq \frac{M_0 \gamma(u_0)}{\sigma_0} \left(1 - E_2\left(\frac{\sigma_0 l}{2}\right)\right).$$

The above relation follows immediately from (3.8) for a sufficiently small $\alpha>0$. This shows that $(\rho e^{-\alpha t}\phi, u_0 e^{-\alpha t})$ is an upper solution. By Theorem 2.2, a global solution (N,u) exists and satisfies

$$0\le N(t,x,\mu)\le \rho e^{-\alpha t}\phi(x,\mu), \qquad 0\le u(t,x)\le u_0 e^{-\alpha t}. \tag{3.12}$$

From the recursion formula (2.10) with $(\underline{N}^{(0)},\underline{u}^{(0)})=(0,0)$ and observing that $f_i(0,0)=0$, $i=1,2$, the first iteration $(\underline{N}^{(1)},\underline{u}^{(1)})$ is given by

$$\begin{aligned}&\underline{N}^{(1)}(t,x,\mu)=\exp(-v\sigma_0 t)N_0(x-v\mu t,\mu),\\ &\underline{u}^{(1)}(t,x)=u_0\exp(-(\beta+M)t).\end{aligned} \tag{3.13}$$

But $(\underline{N}^{(1)},\underline{u}^{(1)})$ is also a lower bound of (N,u); the relation (3.19) follows from (3.12). This completes the proof of the theorem.

The result of Theorem 3.2 implies that under the condition (3.8) the zero steady-state solution is exponentially asymptotically stable. In the following theorem we show that under a similar condition on the same set of physical parameters the solution grows unbounded either as $t\to\infty$ or at a finite time.

THEOREM 3.3. *Let* (H_0), (H_1) *hold for every* $\rho<\infty$. *If*

$$\frac{\sigma_0}{\gamma(u_0)}<\frac{[1-E_2(\sigma_0 l)]}{2} \tag{3.14}$$

then for any $N_0\ge\delta\phi$ *with* $\delta>\beta u_0/\underline{h}(u_0)$ *there exists a constant* $\varepsilon>0$ *such that the system* (2.1)–(2.3) *has a unique solution* (N,u) *which satisfies*

$$N(t,x,\mu)\ge\delta e^{\varepsilon t}\phi, \qquad u(t,x)\ge u_0 e^{\varepsilon t} \qquad (t>0,\ (x,\mu)\in\bar{D}_2) \tag{3.15}$$

for as long as it exists.

Proof. Let M^* be an arbitrarily large constant and define modified functions γ^*, h^* such that $\gamma^*(u)=\gamma(u)$, $h^*(u)=h(u)$ for $0\le u\le M^*$ and γ^*, h^* are uniformly bounded and satisfy the hypotheses (H_0), (H_1) for all $u\ge0$ (for example, $\gamma^*(u)=\gamma(u)$ for $0\le u\le M^*$ and $\gamma^*(u)=\gamma(M^*)$ for $u>M^*$). Then by Theorem 2.2 the modified problem (2.1)–(2.3) (i.e., with γ,h replaced by γ^*,h^*) has a unique solution (N^*,u^*) and satisfies $(\underset{\sim}{N},\underset{\sim}{u})\le(N^*,u^*)\le(\tilde{N},\tilde{u})$ provided that $(\tilde{N},\tilde{u})$, $(\underset{\sim}{N},\underset{\sim}{u})$ are upper and lower solutions of this modified system. We first seek a lower solution in the form $(\underset{\sim}{N},\underset{\sim}{u})=(\delta e^{\varepsilon t}\phi, u_0 e^{\varepsilon t})$. This will be done if it satisfies all the reversed inequalities in (3.3). Indeed, since the boundary and initial requirements are fulfilled it suffices to show that

$$\left[\mu\phi_x+(\sigma_0+v^{-1}\varepsilon)\phi\right]\delta e^{\varepsilon t}\le\frac{\gamma^*(u_0e^{\varepsilon t})}{2}\int_{-1}^{1}\delta e^{\varepsilon t}\phi(x,\mu')\,d\mu',$$

$$(\beta+\varepsilon)u_0e^{\varepsilon t}\le\frac{h^*(u_0e^{\varepsilon t})}{2l}\int_{-1}^{1}\int_0^l\delta e^{\varepsilon t}\phi(x',\mu')\,dx'\,d\mu'.$$

By choosing $\varepsilon>0$ sufficiently small, the second inequality follows from the hypothesis $\beta u_0<\delta\underline{h}(u_0)$. In view of (3.4), (3.5) and the relation $\gamma^*(u_0e^{\varepsilon t})\ge\gamma(u_0)$ the first inequality is also satisfied if

$$(M_0+v^{-1}\varepsilon\phi)\le\frac{\gamma(u_0)}{2}\int_{-1}^{1}\phi(x,\mu')\,d\mu'=\frac{M_0\gamma(u_0)}{2\sigma_0}\omega(x).$$

Since the minimum value of $\omega(x)$ on $[0,l]$ is attained at $x=0$ (or $x=l$) and is given by $\omega(0)=\omega(l)=1-E_2(\sigma_0 l)$ the above inequality holds whenever

$$M_0+v^{-1}\varepsilon\phi\le\frac{M_0\gamma(u_0)}{2\sigma_0}[1-E_2(\sigma_0 l)].$$

But this follows from the condition (3.14) for a sufficiently small $\varepsilon>0$. The above conclusion shows that for a small $\varepsilon>0$, $(\underset{\sim}{N},\underset{\sim}{u})=(\delta e^{\varepsilon t}\phi, u_0 e^{\varepsilon t})$ is a lower solution of the modified problem. Using the uniform boundedness of the modified functions γ^*, h^* the same construction as in the proof of Theorem 3.1 shows that for any constants $\rho_1\geq N_0$, $\rho_2\geq u_0$ and a suitable $\alpha\geq\varepsilon$ the function $(\tilde{N},\tilde{u})=(\rho_1 e^{\alpha t},\rho_2 e^{\alpha t})$ is an upper solution. The above choice of ρ_1, ρ_2, α also ensures that $(\tilde{N},\tilde{u})\geq(\underset{\sim}{N},\underset{\sim}{u})$. It follows from Theorem 2.2 that a unique global solution (N^*,u^*) to the modified problem exists and satisfies

$$\left(\delta e^{\varepsilon t}\phi, u_0 e^{\varepsilon t}\right)\leq\left(N^*,u^*\right)\leq\left(\rho_1 e^{\alpha t},\rho_2 e^{\alpha t}\right).$$

Since (N^*,u^*) is the solution of the original problem whenever $u^*\leq M^*$ and since M^* can be taken arbitrarily large we conclude that either (N^*,u^*) is the solution of the original problem and satisfies (3.15) for all $t\geq 0$ or the solution of (2.1)–(2.3) blows up in a finite time. This completes the proof of the theorem.

Remark 3.1. (a) The results of Theorems 3.1 and 3.2 imply that under the condition (3.8) the zero steady-state solution of (2.1)–(2.2) is asymptotically stable while under the condition (3.14) it is unstable. A stability region and an instability region are given, respectively, by

$$\begin{aligned}\Lambda_s&=\left\{(N_0,u_0)\geq(0,0);\ N_0\leq\rho_0\phi \text{ with } \rho_0<\frac{\beta u_0}{\bar{h}(u_0)}\right\},\\ \Lambda_i&=\left\{(N_0,u_0)\geq(0,0);\ N_0\geq\delta_0 \text{ with } \delta>\frac{\beta u_0}{\underline{h}(u_0)}\right\}.\end{aligned}\tag{3.16}$$

Notice that if $h(u)=h_0$ is independent of u then $\bar{h}(u_0)=\underline{h}(u_0)=h_0$ and the stability and instability regions given by (3.16) become rather sharp. On the other hand, if $\gamma(u)\leq b_1$, $h(u)\leq b_2$ then for any $N_0\geq 0$ there exist u_0, ρ_0 such that $N_0\leq\rho_0<\beta u_0/b_2$. This implies that if (3.8) holds with $\gamma(u_0)$ replaced by b_1 then the solution (N,u) satisfies (3.9) and thus it decays to $(0,0)$ as $t\to\infty$. Hence condition (3.8) improves the stability condition $b_1<\sigma_0$ in Theorem 3.1 when γ, h are uniformly bounded.

(b) The dependence of γ on temperature is reflected in the stability and instability conditions (3.8), (3.14). In the present situation of positive temperature feedback this dependency only involves the initial temperature increment u_0. In the special case where $\gamma(u_0)=\gamma_0$ is independent of u_0 the stability and instability conditions are reduced, respectively, to

$$\sigma_0/\gamma_0>1-E_2\left(\frac{\sigma_0 l}{2}\right)\quad\text{and}\quad\frac{\sigma_0}{\gamma_0}<\frac{[1-E_2(\sigma_0 l)]}{2}.$$

These are exactly the same conditions as for the linear transport equation obtained in [9]. Notice that since $(1-E_2(\sigma_0 l/2))$ is approximately equal to $\frac{1}{2}(1-E_2(\sigma_0 l))$ when $\sigma_0 l$ is sufficiently small, conditions (3.8), (3.14) for stability and instability give a criticality result for small values of $\sigma_0 l$. For large values of $\sigma_0 l$ there is a gap between these two functions whose ratio is always less than 2 for every finite $\sigma_0 l$.

(c) When $u_0=u_0(x)$ depends on x all the conclusions in Theorems 3.2 and 3.3 remain true if $\gamma(u_0)$ in the conditions (3.8) and (3.14) is replaced by $\gamma(\bar{u}_0)$ and $\gamma(\underline{u}_0)$, respectively, where $\bar{u}_0$, $\underline{u}_0$ denote the least upper bound and greatest lower bound of u_0 on $[0,l]$. In either case, whether u_0 is constant or not, if the temperature increment u_0 from the coolant temperature T_c is small, the feedback effect on the stability or instability of the steady-state $(0,T_c)$ differs little from the linear system.

(d) The slab length l also plays an important role in the stability or instability of the system. In fact, given any σ_0, γ, u_0, one can determine a slab length l, from (3.8) and l_2 from (3.14) so that the system (2.1)–(2.3) is stable for all $l \le l_1$ and is unstable for all $l \ge l_2$.

When the functions γ, h are given by (1.5) with $m \ge 1$, $n \ge 1$ all the conditions in (H_0), (H_1) are satisfied. In terms of the initial temperature $T_0 \ge T_c$ the functions $\gamma(u_0)$, $h(u_0)$ become

$$\gamma(u_0) = \gamma_1(T_0) = \gamma_0 + a_1 \ln\left(\frac{T_0}{T_c}\right) \qquad \text{for } m = 1,$$
$$\gamma(u_0) = \gamma_m(T_0) = \gamma_0 - a_1(m-1)^{-1}\left(T_0^{-(m-1)} - T_c^{-(m-1)}\right) \qquad \text{for } m \ge 1 \tag{3.17}$$

and a similar expression of $h(u_0)$. As a direct consequence of Theorem 3.2 and 3.3 we have the following conclusion for the system (1.1)–(1.3) when γ, h are given by (1.5).

COROLLARY. *Let* $\gamma = \gamma_1(T)$, $h = h_1(T)$ *be given by* (1.5) *with* $m = n = 1$ *and let* $T_0 \ge T_c$. *Then the steady-state* $(0, T_c)$ *of* (1.1), (1.2) *is asymptotically stable if*

$$\sigma_0\left[\gamma_0 + a_1 \ln\left(\frac{T_0}{T_c}\right)\right]^{-1} > \left[1 - E_2\left(\frac{\sigma_0 l}{2}\right)\right] \tag{3.18}$$

and it is unstable if

$$\sigma_0\left[\gamma_0 + a_1 \ln\left(\frac{T_0}{T_c}\right)\right]^{-1} < \frac{[1 - E_2(\sigma_0 l)]}{2}. \tag{3.19}$$

Similarly, when $\gamma = \gamma_m(T)$, $h = h_m(T)$ *with* $m > 1$, $n > 1$, *the stability and instability conditions for* $(0, T_c)$ *are given, respectively, by*

$$\sigma_0\left[\gamma_0 - a_1(m-1)^{-1}\left(T_0^{-(m-1)} - T_c^{-(m-1)}\right)\right]^{-1} > \left[1 - E_2\left(\frac{\sigma_0 l}{2}\right)\right] \tag{3.20}$$

and

$$\sigma_0\left[\gamma_0 - a_1(m-1)^{-1}\left(T_0^{-(m-1)} - T_c^{-(m-1)}\right)\right]^{-1} < \frac{[1 - E_2(\sigma_0 l)]}{2}. \tag{3.21}$$

In each case, a stability and an instability region are given by (3.16) *with* $u_0 = T_0 - T_c$.

4. Blowing-up property of the solution. It is seen from Theorem 3.3 that under the condition (3.14) the solution (N, u) of (2.1)–(2.3) grows unbound either at infinity or in a finite time. Clearly for uniformly bounded functions γ, h, global solutions always exist and thus (N, u) cannot grow faster than exponential order. A mathematically interesting question is that for nonuniformly bounded functions γ, h whether the solution can grow unbounded in finite time and whether it can grow faster than an exponential order if a global solution does exist. In particular, it is interesting to know these possibilities for the type of functions γ, h given by (1.5) when $m \le 1$ and $n \le 1$. The aim of this section is to investigate the existence and nonexistence of a global solution as well as the growth or decay property of the solution. We first show that if γ, h satisfy the condition

$$\gamma(u) \ge \gamma_0 + b_1 u^{\nu_1}$$
$$h(u) \ge h_0 + b_2 u^{\nu_2} \qquad (u \ge 0) \tag{4.1}$$

for some constants $\gamma_0 \geq 0$, $h_0 \geq 0$, $b_i > 0$, $\nu_i > 0$, no matter how small b_i, ν_i may be $(i = 1, 2)$, then for certain classes of initial functions the corresponding solutions blow up in finite time. In the special model (1.5), condition (4.1) corresponds to the case $m < 1$, $n < 1$. In order to give more specific information about this class of initial functions, it is convenient to use the following notations:

$$\begin{aligned} &\nu = \min\{\nu_1, \nu_2\}, \quad c_0 = \frac{(1 - E_2(\sigma_0 l))}{2}, \quad B = b_2 (v b_1 c_0)^{-1}, \\ &B^* = b_2^{\nu} (v b_1 c_0)^{1-\nu}, \quad \beta^* = \max\{\beta - h_0 B^{-1}, v(\sigma_0 - c_0 \gamma_0), 0\}. \end{aligned} \tag{4.2}$$

THEOREM 4.1. *Let γ, h satisfy the condition* (4.1) *and the hypotheses* (H_0), (H_1) *for every $\rho < \infty$, and let $\delta > (\beta^*/B^*)^{1/\nu}$. Then for any $(N_0, u_0) \geq (\delta\phi, \delta B)$ there exists a finite $t^* > 0$ such that a unique (local) solution (N, u) exists on $[0, t^*) \times [0, l] \times [-1, 1]$ and satisfies either*

$$\lim_{t \to t^*} \left[\max_{(x,\mu) \in D_2} N(t, x, \mu) \right] = \infty \quad \text{or} \quad \lim_{t \to t^*} \left[\max_{0 \leq x \leq l} u(t, x) \right] = \infty \tag{4.3}$$

(or both). Moreover, $t^* \leq \nu B^* \delta^{\nu}$ when $\beta^* = 0$ and

$$t^* \leq (\nu \beta^*)^{-1} \ln\left[B^* \delta^{\nu} (B^* \delta^{\nu} - \beta^*)^{-1} \right] \quad \text{when } \beta^* > 0. \tag{4.4}$$

Proof. We first seek a lower solution in the form $(\underset{\sim}{N}, \underset{\sim}{u}) = (p(t)\phi, q(t))$, where p, q are some positive functions with $p(0) = \delta$, $q(0) = \delta B$. To this end, it suffices to find p, q such that $(\underset{\sim}{N}, \underset{\sim}{u})$ satisfies all the reversed inequalities in (3.3). Since the boundary and initial requirements are fulfilled we only need to choose $(p, q) > (0, 0)$ such that

$$\begin{aligned} &v^{-1} p' \phi + (\mu \phi_x + \sigma_0 \phi) p \leq \frac{\gamma(q) p}{2} \int_{-1}^{1} \phi(x, \mu')\, d\mu', \\ &q' + \beta q \leq \frac{h(q) p}{2l} \int_{-1}^{1} \int_{0}^{l} \phi(x', \mu')\, dx'\, d\mu' = h(q) p. \end{aligned} \tag{4.5}$$

Now from the expression for ϕ in (3.5),

$$\phi(x, \mu) \leq \frac{M_0}{\sigma_0} \quad \text{and} \quad \int_{-1}^{1} \phi(x, \mu')\, d\mu' \geq \frac{M_0}{\sigma_0} (1 - E_2(\sigma_0 l)).$$

This relation together with (3.4) and the hypothesis (4.1) imply that (4.5) will be satisfied if $p' \geq 0$ and

$$\begin{aligned} &\frac{M_0}{\sigma_0 v} p' + M_0 p \leq \frac{M_0}{2\sigma_0} (1 - E_2(\sigma_0 l))(\gamma_0 + b_1 q^{\nu_1}) p, \\ &q' + \beta q \leq (h_0 + b_2 q^{\nu_2}) p. \end{aligned} \tag{4.6}$$

Choose $q = Bp$, where B is given by (4.2). Then by using the notation in (4.2) the above relation holds when

$$\begin{aligned} &p' + v(\sigma_0 - c_0 \gamma_0) p \leq v b_1 c_0 B^{\nu} p^{1+\nu}, \\ &p' + (\beta - h_0 B^{-1}) p \leq b_2 B^{\nu - 1} p^{1+\nu}. \end{aligned} \tag{4.7}$$

Since $v b_1 c_0 B^{\nu} = b_2 B^{\nu - 1} = B^*$, both inequalities in (4.7) hold if p satisfies

$$p' + \beta^* p \leq B^* p^{1+\nu}$$

where β^*, B^* are given by (4.2). It follows that for $\beta^*>0$ p can be chosen as

$$p(t)=\delta\left[\frac{B^*\delta^\nu}{\beta^*}-\left(\frac{B^*\delta^\nu}{\beta^*}-1\right)e^{\nu\beta^* t}\right]^{-1/\nu} \qquad (t\in[0,t_0)), \tag{4.8}$$

where $p(0)=\delta>(\beta^*/B^*)^{1/\nu}$ and t_0 is given by the right side of (4.4). The requirement of $\delta^\nu>\beta^*/B^*$ ensures that p' and p are both positive on $[0,t_0)$ and $p(t)\to\infty$ as $t\to t_0$. When $\beta^*=0$ it suffices to choose $p(t)=\delta(1-\nu B^*\delta^\nu t)^{-1/\nu}$ for $t<(\nu B^*\delta^\nu)^{-1}$. With this choice of p, $(\underset{\sim}{N},\underset{\sim}{u})=(p\phi,Bp)$ is a lower solution in the domain $Q_1=(0,t_1]\times D_2$ for every $t_1<t_0$. To show the blowing-up property (4.3) we follow a similar argument as in [11] where a diffusion system was considered. Assume, by contradiction, that the solution of (2.1)–(2.3) were bounded on $\bar{Q}_0\equiv[0,t_0]\times\bar{D}_2$ (say, by A). Then there exists $t_1<t_0$ such that either $\max(p\phi)\geq A+1$ or $\max(Bp)\geq A+1$ (or both) in the region $\bar{Q}_1$. Choose $M^*\geq A+1$ sufficiently large and define modified functions γ^*, h^* as in the proof of Theorem 3.3. Then $(p\phi,Bp)$ remains a lower solution of the modified system (2.1)–(2.3) in the domain Q_1. By choosing ρ,α sufficiently large the function $(\tilde{N},\tilde{u})=(\rho e^{\alpha t},\rho e^{\alpha t})$ is an upper solution of the modified problem in Q_1 and $(\tilde{N},\tilde{u})\geq(\underset{\sim}{N},\underset{\sim}{u})$ on $\bar{Q}_1$. Hence the modified problem has a unique solution (N^*,u^*) such that $(N^*,u^*)\geq(p\phi,Bp)$ on $\bar{Q}_1$. This relation implies that for some $t_2\leq t_1$, $(N^*,u^*)\leq(M^*,M^*)$ and either $\max N^*=A+1$ or $\max u^*=A+1$ in the region $\bar{Q}_2=[0,t_2]\times\bar{D}_2$. Hence (N^*,u^*) is a solution of the original problem in Q_2 with either $\max N^*=A+1$ or $\max u^*=A+1$ on $\bar{Q}_2$. This contradicts the assumption that the solution be bounded by A on $\bar{Q}_0$. Therefore there must exist some $t^*\leq t_0$ such that (4.3) holds. This completes the proof of the theorem.

When the functions γ,h are given by

$$\gamma(u)=\gamma_0+b_1u^{\nu_1}, \qquad h(u)=h_0+b_2u^{\nu_2} \qquad (\nu_i\geq 1) \tag{4.9}$$

all the conditions in (H_0), (H_1) and (4.1) are satisfied for every $\rho<\infty$. In view of Theorems 3.2 and 4.1 we have

COROLLARY. *Let γ,h be given by* (4.9) *and let* (3.8) *hold with $\gamma(u_0)=\gamma_0+b_1u_0^{\nu_1}$. Then for any $(N_0,u_0)\geq(0,0)$ with $N_0\leq\rho_0\phi$ and $\rho_0<\beta u_0(h_0+b_2u_0^{\nu_2})^{-1}$, a unique global solution to* (2.1)–(2.3) *exists and converges to* $(0,0)$ *as $t\to\infty$; while for $(N_0,u_0)\geq(\delta\phi,\delta B)$ with $\delta>(\beta^*/B^*)^{1/\nu}$ the corresponding solution blows up in finite time.*

Remark 4.1. The result of Theorem 4.1 remains true when the first condition in (4.1) is satisfied only for $u\geq\delta\phi$ and the second condition is satisfied for $u\geq B\delta$ (rather than for all $u\geq 0$). The blowing-up property of the solution in the corollary to Theorem 4.1 remains true when condition (4.9) holds only for $\nu_i>0$, no matter how small ν_i may be. (Note that in this situation, Theorem 2.1 ensures the existence of a solution but not uniqueness.) In any case a "strong instability region" is given by

$$\Lambda=\left\{(N_0,u_0);\ N_0\geq\delta\rho,u_0\geq\delta B \text{ with } \delta>(\beta^*/B^*)^{1/\nu}\right\},$$

where B,B^*, β^*, ν are defined in (4.2).

The blowing-up property of the solution obtained in Theorem 4.1 is based on the condition in (4.1) which is not physically realistic since it corresponds to $m<1$, $n<1$ in the model (1.5). Nevertheless it is an interesting mathematical problem since this conclusion is in sharp contrast to the case $m>1$, $n>1$ for which global solution always exists for every nonnegative (N_0,u_0). An immediate question is that when $m=n=1$ whether global solutions can exist for all nonnegative (N_0,u_0). Since in this model,

$$\gamma(u)=\gamma_0+a_1\ln\left(1+\frac{u}{T_c}\right), \qquad h(u)=h_0+a_2\ln\left(1+\frac{u}{T_c}\right) \tag{4.10}$$

which grows unbounded as $u \to \infty$ but not as fast as u^{ν} one cannot draw a conclusion from either Theorem 3.2 or Theorem 4.1. It turns out that for this model, global solutions do exist for all nonnegative (N_0, u_0) but as $t \to \infty$ they may grow to infinity at a rate faster than exponential order. In order to exhibit this type of growth property of the solution more explicitly we set

$$
\begin{aligned}
c_1 &= \max\{a_1^{-1}(\gamma_0 - \sigma_0), a_2^{-1}h_0 - (va_1)^{-1}\beta\}, \\
c_2 &= \min\{a_1^{-1}(\gamma_0 - \sigma_0/c_0), a_2^{-1}h_0 - (va_1c_0)^{-1}\beta\},
\end{aligned} \tag{4.11}
$$

where c_0 is defined in (4.2). Notice that the constants c_1, c_2 may be positive as well as negative, depending on the magnitude of the various constants.

THEOREM 4.2. *Let γ, h be given by* (4.10) *and let*

$$
\begin{aligned}
q_1(t) &= T_c[\exp(C_1 \exp(va_1 t) - c_1) - 1], \\
q_2(t) &= T_c[\exp(C_2 \exp(va_1c_0 t) - c_2) - 1],
\end{aligned} \tag{4.12}
$$

where C_1, C_2 are some positive constants. Then for any $(N_0, u_0) \geq (0,0)$ there exists a constant $C_1 \geq c_1$ such that a unique global solution (N, u) exists and satisfies

$$
\begin{aligned}
&N_0(x - v\mu t, \mu)e^{-v\sigma_0 t} \leq N(t,x,\mu) \leq (a_2^{-1}va_1)q_1(t), \\
&u_0(x)e^{-(\beta+M)t} \leq u(t,x) \leq q_1(t).
\end{aligned} \tag{4.13}
$$

Moreover, if $(N_0, u_0) \geq (\delta_1\phi, \delta_2)$ for some $\delta_2 \geq (e^{-c_2} - 1)$, $\delta_1 \geq (a_2^{-1}va_1c_0)\delta_2$ then there exists a positive constant $C_2 \leq C_1$ such that the solution (N, u) satisfies

$$
\begin{aligned}
&(a_2^{-1}va_1c_0)q_2(t) \leq N(t,x,\mu) \leq (a_2^{-1}va_1)q_1(t), \\
&q_2(t) \leq u(t,x) \leq q_1(t).
\end{aligned} \tag{4.14}
$$

In particular, if $c_2 \geq 0$ then for every positive (N_0, u_0), the corresponding solution grows to infinity in the order no less than $\exp(\exp(va_1c_0 t))$ *and no more than* $\exp(\exp(va_1 t))$.

Proof. For the relation (4.13) we seek an upper and a lower solution in the form $(\tilde{N}, \tilde{u}) = (p(t), q(t))$, $(\underset{\sim}{N}, \underset{\sim}{u}) = (0,0)$ where p, q are some positive functions to be determined. Since $(0,0)$ is a lower solution it suffices to find $(p,q) \geq (0,0)$ such that $(p(0), q(0)) \geq (N_0, u_0)$ and

$$
v^{-1}p' + \sigma_0 p \geq \gamma(q)p, \qquad q' + \beta q \geq h(q)p.
$$

Choose $p = (a_2^{-1}va_1)q$, where $q(0) \geq u_0$ and $(a_2^{-1}va_1)q(0) \geq N_0$. Then by (4.10) the requirement on (p,q) becomes

$$
\begin{aligned}
&q' + v\sigma_0 q \geq v[\gamma_0 + a_1 \ln(1 + a_0 q)]q, \\
&q' + \beta q \geq (a_2^{-1}va_1)[h_0 + a_2 \ln(1 + a_0 q)]q,
\end{aligned} \tag{4.15}
$$

where $a_0 = T_c^{-1}$. Using the notations in (4.11), both inequalities in (4.15) hold if

$$
q' \geq va_1[c_1 + \ln(1 + a_0 q)]q. \tag{4.16}
$$

Let $r(t) = 1 + a_0 q(t)$. Then the above inequality becomes

$$
r'(t) \geq va_1[(c_1 + \ln r)r - (c_1 + \ln r)]. \tag{4.17}
$$

Now if $c_1 \geq 0$ then from $c_1 + \ln r(0) = c_1 + \ln(1 + a_0 q(0)) \geq 0$ it suffices to find $r(t)$ such that

$$
r'(t) \geq va_1(c_1 + \ln r)r, r(0) = 1 + q_0 q(0).
$$

This leads to the choice of

$$r(t)=\exp\left[(c_1+\ln r(0))\exp(va_1t)-c_1\right]. \tag{4.18}$$

Hence a desired function $q(t)$ may be taken as the function $q_1(t)$ in (4.12) with $C_1=c_1+\ln(1+a_0q(0))$. With this choice of q_1 the function $(p,q)=(a_2^{-1}va_1q_1,q_1)$ is an upper solution. It follows from Theorem 2.2 that a unique global solution (N,u) to (2.1)–(2.3) exists and

$$0\le N(t,x,\mu)\le\left(a_2^{-1}va_1\right)q_1(t),\qquad 0\le u(t,x)\le q_1(t). \tag{4.19}$$

Since the function $(\underset{\sim}{N},\underset{\sim}{u})$ given by (3.13) is a lower bound of the solution we conclude that (N,u) satisfies the relation (4.13). In the case of $c_1<0$ the relation (4.17) is still satisfied by the function $r(t)$ in (3.18) when $q(0)$ is chosen such that $\ln(1+a_0q(0))\ge -c_1$. In either case, whether $c_1\ge 0$ or $c_1<0$, the conclusion in (4.13) holds for every nonnegative (N_0,u_0) (where C_1 depends on (N_0,u_0)).

To show the relation (4.14) when $(N_0,u_0)\ge(\delta_1\phi,\delta_2)$ we seek a different lower solution in the form $(\underset{\sim}{N},\underset{\sim}{u})=(p_1\phi,p_2)$ where p_1, p_2 are some positive functions with $p_1(0)=\delta_1$, $p_2(0)=\delta_2$. Following the same argument as in the proof of Theorem 4.1, $(p_1\phi,p_2)$ is a lower solution if $p_1'\ge 0$ and

$$\begin{aligned}&\frac{M_0}{\sigma_0 v}p_1'+M_0p_1\le\frac{M_0}{2\sigma_0}\left(1-E_2(\sigma_0 l)\right)\left[\gamma_0+a_1\ln(1+a_0p_2)\right]p_1,\\ &p_2'+\beta p_2\le\left[h_0+a_2\ln(1+a_0p_2)\right]p_1\end{aligned} \tag{4.20}$$

(see (4.6)). Let $p_1=a_2^{-1}(va_1c_0)p_2$, where c_0 is given by (4.2). Then a simple calculation shows that both inequalities in (4.20) are verified if p_2 satisfies the relation

$$p_2'\le va_1c_0\left[c_2+\ln(1+a_0p_2)\right]p_2, \tag{4.21}$$

where c_2 is defined by (4.11). Since the above inequality has exactly the same form as in (4.16) except with c_1, va_1 replaced by c_2 and va_1c_0, respectively, it is satisfied by the function $q_2(t)$ in (4.12) provided that $C_2=c_2+\ln(1+p_2(0))\ge 0$. The condition $C_2\ge 0$ is insured by the hypothesis $p_2(0)=\delta_2\ge e^{-c_2}-1$. Since $p_1(0)\phi=(a_2^{-1}va_1c_0)\delta_2\phi\le\delta_1\phi\le N_0$ we see that $(p_1\phi,p_2)=((a_2^{-1}va_1c_0)q_2\phi,q_2)$ is a lower solution. It is easily seen from $c_0<1$ that $q_2\le q_1$ when $C_2\le C_1$. By choosing C_1 sufficiently large (or equivalently $q(0)$ large), if necessary, we also have $q_2\phi\le q_1$. This choice of C_1 implies that $(p_2\phi,q_2)\le(p,q)$, that is, $(\underset{\sim}{N},\underset{\sim}{u})\le(\tilde{N},\tilde{u})$. The relation (4.14) follows from Theorem 2.2. Finally, if $c_2\ge 0$ then $c_2+\ln(1+p_2(0))\ge 0$ for every $p_2(0)\ge 0$. In this situation, the relation (4.14) holds for every positive (N_0,u_0) (with a corresponding $C_1\ge C_2\ge 0$). The growth property of the solution as stated in the theorem follows from the relation (4.12), (4.14). The proof of the theorem is completed.

REFERENCES

[1] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, Van Nostrand-Reinhold, New York, 1970.

[2] A. Belleni-Morante, *Neutron transport with temperature feedback*, Nucl. Sci. Eng., 59 (1976), pp. 56–58.

[3] G. Busoni, V. Capasso and A. Belleni-Morante, *Global solution for a problem of neutron transport with temperature feedback*, in Nonlinear Systems and Applications, V. Lakshmikantham, ed., Academic Press, New York, 1977.

[4] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison-Wesley, Reading, MA, 1967.

[5] John Graham, *Faster Reactor Safety*, Academic Press, New York, 1971.

[6] W. E. KASTENBERG AND P. L. CHAMBRÉ, *On the stability of nonlinear space-dependent reactor kinetics*, Nucl. Sci. Eng., 31 (1968), pp. 67–79.
[7] P. DE MOTTONI AND A. TESEI, *Asymptotic stability results for a system of quasilinear parabolic equations*, J. Applicable Anal., 9 (1979), pp. 7–21.
[8] D. H. NGUYEN, *Stability of nuclear reactors with changes in eigenvalue*, Nucl. Sci. Eng., 50 (1973), pp. 370–381.
[9] C. V. PAO, *Asymptotic behavior of the solution for the time-dependent neutron transport problem*, J. Integral Equations, 1 (1979), pp. 131–152.
[10] _____, *Positive solutions and criticality of the linear and some nonlinear neutron transport problems*, SIAM J. Appl. Math., 32 (1977), pp. 164–176.
[11] _____, *Bifurcation analysis on a nonlinear diffusion system in reactor dynamics*, J. Applicable Anal., 9 (1979), pp. 107–119.
[12] D. B. REISTER AND P. L. CHAMBRÉ, *Solution bounds for nonlinear space-time reactor problems*, Nucl. Sci. Eng., 48 (1972), pp. 211–218.
[13] W. M. STACEY, JR., *Space-Time Nuclear Reactor Kinetics*, Academic Press, New York, 1969.
[14] M. M. R. WILLIAMS, *Mathematical Methods in Particle Theory*, John Wiley, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0015 $01.25/0

CONTINUOUS SELECTIONS AND APPROXIMATE SELECTION FOR SET-VALUED MAPPINGS AND APPLICATIONS TO METRIC PROJECTIONS*

FRANK DEUTSCH† AND PETAR KENDEROV‡

Abstract. Two new continuity properties for set-valued mappings are defined which are weaker than lower semicontinuity. One of these properties characterizes when approximate selections exist. A few selection theorems characterized by the other property are established. Some applications are made to set-valued metric projections.

Key words. set-valued mapping, continuous selection, continuous approximate selection, metric projection, continuous selections for metric projections

1. Introduction. The most important mapping which arises in approximation theory is the set-valued metric projection or proximity map. This is the mapping which associates with each element of a normed linear space its set of nearest points ("best approximations") from a prescribed subset. In particular, there has been substantial interest in determining conditions which insure that the metric projection onto a finite-dimensional subspace has a continuous selection. The well-known Michael selection theorem (see below) states that lower semicontinuity of the metric projection is sufficient. Unfortunately, lower semicontinuity is not necessary and, in fact, is often absent in any given problem. This has forced many researchers to employ ad hoc techniques to establish the existence (or nonexistence) of continuous selections for the metric projection (see e.g. [2], [3], [10], [11], [13], [14], [16], [17], [20], [21], [22]).

The main motivation for this paper was a desire to determine a *continuity property* of the metric projection which *characterizes* when a continuous selection exists. Moreover, since the key step in the proof of the Michael selection theorem was the construction of "continuous ε-approximate" selections, we sought a continuity criterion which characterized this latter condition as well. Our results can be very briefly summarized as follows: We were successful on the latter problem and were partially successful on the former. More precisely, we proved that the existence of continuous ε-approximate selections for each $\varepsilon>0$ is equivalent to "almost lower semicontinuity" (Theorem 2.4). And, in certain cases of interest, the existence of a continuous selection is equivalent to "2-lower semicontinuity" (Theorems 2.7 and 2.9 and Corollaries 3.3 and 3.5).

Since our results are valid for rather general set-valued mappings (not necessarily metric projections) and have independent interest, we have formulated them in a general setting in §2 and then applied these results to metric projections in §3.

An announcement, without proofs, of some of the results of this paper involving metric projections was made in [6]. (In [6], "n-lower semicontinuity" was called "n-continuity".)

*Received by the editors July 11, 1981, and in revised form February 1, 1982.

†Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802. The work of this author was partially supported by the Fachbereich Mathematik der Universität Frankfurt, where he spent the academic year 1978–79 as visiting professor.

‡Bulgarian Academy of Sciences, Institute of Mathematics, 1000 Sofia, Bulgaria. The work of this author was partially supported by the Alexander von Humboldt Foundation, where he spent the academic year 1978–79 as Research Scientist.

2. Almost lower semicontinuity and *n*-lower semicontinuity. Let X be a topological space, Y a metric space with metric d, and 2^Y the collection of all nonempty subsets of Y. Let $F: X \to 2^Y$. That is, F is a function on X whose images are nonempty subsets of Y. The *ε-neighborhood* of a nonempty set A in Y is given by

$$B_\varepsilon(A) := \{y \in Y \mid d(y,A) < \varepsilon\},$$

where

$$d(y,A) = \inf\{d(y,a) \mid a \in A\}.$$

DEFINITIONS 2.1. F is called *almost lower semicontinuous* (a.l.s.c.) (resp. *n-lower semicontinuous* (n-l.s.c.)) at $x_0 \in X$ if for each $\varepsilon > 0$, there exists a neighborhood U of x_0 such that

$$\bigcap_{x \in U} B_\varepsilon(F(x)) \neq \varnothing$$

(resp. $\bigcap_1^n B_\varepsilon(F(x_i)) \neq \varnothing$ for each choice of n points $x_1, x_2, \cdots, x_n$ in U). F is called *almost lower semicontinuous* (a.l.s.c.) (resp. *n-lower semicontinuous* (n-l.s.c.)) if F is a.l.s.c. (resp. n-l.s.c.) at each point of X.

A *continuous selection* (resp. *continuous ε-approximate selection*) for F is a continuous function $f: X \to Y$ such that $f(x) \in F(x)$ (resp. $f(x) \in B_\varepsilon(F(x))$) for each $x \in X$. ("Approximate selections" have been studied by Cellina [4], [5] and Reich [18], but their definition differs significantly from ours.)

It is useful for comparison purposes to mention here the celebrated selection (resp. approximate selection) theorem of Michael [12]: *Let X be a paracompact space, Y a Banach space (resp. normed linear space), and suppose $F: X \in 2^Y$ has closed convex (resp. convex) images. If F is lower semicontinuous, then F has a continuous selection (resp. a continuous ε-approximate selection for each $\varepsilon > 0$).*

Recall that F is *lower semicontinuous* at x_0 if for any open set V in Y with $F(x_0) \cap V \neq \varnothing$, there exists a neighborhood U of x_0 such that $F(x) \cap V \neq \varnothing$ for all $x \in U$. It is easy to give examples which show that lower semicontinuity of F is *not* necessary for F to admit either a continuous selection or a continuous ε-approximate selection (see the examples following Lemma 2.2). To the best of our knowledge, the result of Michael's is the only one previously known about continuous ε-approximate selections.

Some easy consequences of the definitions are recorded in the following lemma.

LEMMA 2.2. (1) *F is always* 1-l.s.c..

(2) *If F is n-l.s.c., then F is k-l.s.c. for every $k \leq n$.*

(3) *If $n \geq 2$ and F is singleton-valued, then F is* n-l.s.c. *or* a.l.s.c. *if and only if F is continuous (in the usual sense).*

(4) *If F is* a.l.s.c. *(resp. n-l.s.c.), then the mapping $x \mapsto \overline{F(x)}$ is also* a.l.s.c. *(resp. n-l.s.c.) (since $B_\varepsilon(\overline{F(x)}) = B_\varepsilon(F(x))$).*

(5) *Every continuous selection is a continuous ε-approximate selection.*

It is worth noticing that a mapping F could have a continuous ε-approximate selection for each $\varepsilon > 0$, yet fail to have a continuous selection. For example, define $F: \mathbb{R} \to 2^{\mathbb{R}}$ by $F(x) = [-1, 0)$ if x is rational and $F(x) = (0, 1]$ if x is irrational. Then for any $\varepsilon > 0$, the function $f = f_\varepsilon \equiv 0$ is a continuous ε-approximate selection for F, but F obviously has no continuous selection. Furthermore, F is not l.s.c. This shows that lower semicontinuity is *not* necessary in Michael's approximate selection theorem stated above. However, F *is* a.l.s.c. and this is the reason why F admits continuous ε-approximate selections for every $\varepsilon > 0$ (see Theorem 2.4 below).

By slightly modifying the above example, we can show that lower semicontinuity is *not* necessary in Michael's selection theorem. Indeed, define $F: \mathbb{R} \to 2^{\mathbb{R}}$ by $F(x)=[-1,0]$ if x is rational and $F(x)=[0,1]$ if x is irrational. Then F is *not* l.s.c. but admits the continuous selection $f \equiv 0$. The reason F admits a continuous selection is because F is 2-l.s.c. (see Theorem 2.7 below).

The next lemma exhibits the hierarchy of these continuity properties and shows, in particular, the almost lower semicontinuity and n-lower semicontinuity are weaker than either lower semicontinuity or the admission of a continuous selection.

LEMMA 2.3. *Consider the following statements.*

(1) *F is* l.s.c..

(2) *F has a continuous selection.*

(3) *F has continuous ε-approximate selections for each $\varepsilon>0$.*

(4) *F is* a.l.s.c..

(5) *F is n-*l.s.c. *for every n.*

(6) *F is n-*l.s.c. *for some $n \geq 2$.*

(7) *F is* 2-l.s.c.

Then (2)$\Rightarrow$(3)$\Rightarrow$(4)$\Rightarrow$(5)$\Rightarrow$(6)$\Rightarrow$(7) *and* (1)$\Rightarrow$(4).

Proof. The implications (2)$\Rightarrow$(3) and (4)$\Rightarrow$(5)$\Rightarrow$(6)$\Rightarrow$(7) are obvious.

(3)$\Rightarrow$(4). Assume (3) holds, $x_0 \in X$, and $\varepsilon>0$. Let $f=f_\varepsilon$ be a continuous $\frac{\varepsilon}{2}$-approximate selection for F. Choose a neighborhood U of x_0 such that $d(f(x_0), f(x))<\frac{\varepsilon}{2}$ for all $x \in U$. Hence $f(x_0) \in B_{\varepsilon/2}(f(x)) \subset B_\varepsilon(F(x))$ for all $x \in U$. Thus $\bigcap_{x \in U} B_\varepsilon(F(x)) \neq \varnothing$ and F is a.l.s.c. at x_0.

(1)$\Rightarrow$(4). Assume F is l.s.c., $x_0 \in X$, and $\varepsilon>0$. Choose any $y_0 \in F(x_0)$. Obviously, $F(x_0) \cap B_\varepsilon(y_0) \neq \varnothing$ so there exists a neighborhood U of x_0 such that $F(x) \cap B_\varepsilon(y_0) \neq \varnothing$ for all $x \in U$. In particular, $y_0 \in B_\varepsilon(F(x))$ for all $x \in U$ so $\bigcap_{x \in U} B_\varepsilon(F(x)) \neq \varnothing$ and F is a.l.s.c. at x_0. □

Our first theorem characterizes those mappings which have continuous ε-approximate selections for every ε.

THEOREM 2.4. *Let X be a paracompact space (e.g. a metric space) and let Y be a normed linear space. Let $F: X \to 2^Y$ have convex images. Then F is almost lower semicontinuous if and only if for each $\varepsilon>0$, F has a continuous ε-approximate selection.*

Proof. Sufficiency follows from Lemma 2.3.

Necessity. Suppose F is a.l.s.c. and $\varepsilon>0$. For each $x_0 \in X$ there exists a neighborhood $U(x_0)$ of x_0 such that

$$\bigcap \{B_\varepsilon(F(x)) \mid x \in U(x_0)\} \neq \varnothing .$$

Since X is paracompact, the open cover $\{U(x) \mid x \in X\}$ of X has a locally finite refinement $\{V_i \mid i \in I\}$. For each $i \in I$, choose $x_i \in X$ such that $V_i \subset U(x_i)$. Using paracompactness, we can choose a partition of unity $\{p_i \mid i \in I\}$ subordinate to $\{V_i \mid i \in I\}$. That is, each function $p_i: X \to [0,1]$ is continuous, $\Sigma_{i \in I} p_i(x)=1$ for all $x \in X$, and $p_i=0$ off V_i. For each $i \in I$, choose $y_i \in \bigcap \{B_\varepsilon(F(x)) \mid x \in V_i\}$ and set

$$f(x)=\sum_{i \in I} p_i(x) y_i, \qquad x \in X.$$

Given any $x \in X$, there is a neighborhood of x which intersects only finitely many of the V_i so $x \in V_i$ for only a finite set of indices $I(x)$ in I. Thus f is well-defined, continuous, and has range in Y. Further, $y_i \in B_\varepsilon(F(x))$ for all $i \in I(x)$ implies that

$$f(x)=\sum_{i \in I(x)} p_i(x) y_i \in \operatorname{co}(B_\varepsilon(F(x)))=B_\varepsilon(F(x)).$$

Thus f is a continuous ε-approximate selection for F. □

The method of proof of the necessity is similar to that used by Michael [12]. Note that Theorem 2.4 generalizes the approximate selection theorem of Michael stated above.

With an additional restriction on the images of F, we can also relate n-lower semicontinuity to the admission of continuous ε-approximate selections.

THEOREM 2.5. *Let X be a paracompact space, Y an n-dimensional normed linear space, and suppose the mapping $F: X \to 2^Y$ has closed, bounded, and convex images. Then F is $(n+1)$-lower semicontinuous if and only if for each $\varepsilon>0$, F has a continuous ε-approximate selection.*

Proof. Sufficiency follows from Lemma 2.3.

Necessity. Suppose F is $(n+1)$-l.s.c., $x_0 \in X$, and $\varepsilon>0$. Then there exists a neighborhood U of x_0 such that

$$(*) \qquad \bigcap_{i=1}^{n+1} B_{\varepsilon/2}(F(x_i)) \neq \varnothing$$

for each choice of $n+1$ points in U. Each set $\overline{B_{\varepsilon/2}(F(x))}$ is compact convex and is contained in $B_\varepsilon(F(x))$. By (*), the collection of compact convex sets $\{\overline{B_{\varepsilon/2}(F(x))} \mid x \in U\}$ has the property that the intersection of each $n+1$ of them has a nonempty intersection. By Helly's theorem [19, p. 191] the intersection of the whole collection is nonempty. Hence

$$\varnothing \neq \bigcap \{\overline{B_{\varepsilon/2}(F(x))} \mid x \in U\} \subset \bigcap \{B_\varepsilon(F(x)) \mid x \in U\}.$$

This shows that F is a.l.s.c. at x_0. The result now follows by Theorem 2.4. □

For the special case $n=1$, each of the statements of Theorem 2.5 is equivalent to the condition that F have a continuous selection. Before proving this, it is convenient to isolate a key step of the proof.

LEMMA 2.6. *Let Y be a 1-dimensional normed linear space, and suppose the mapping $F: X \to 2^Y$ is 2-lower semicontinuous and has closed, bounded, and convex images. Let $r>0$ and let f be a continuous r-approximate selection for F. Then the mapping $G: X \to 2^Y$ defined by*

$$G(x) = \overline{F(x) \cap B_r(f(x))}, \qquad x \in X$$

is 2-lower semicontinuous and has closed, bounded, and convex images.

Proof. Note first that for each $x \in X$,

$$G(x) \supset F(x) \cap B_r(f(x)) \neq \varnothing$$

since $f(x) \in B_r(F(x))$. Thus $G(x) \in 2^Y$ and G has (closed) bounded and convex images since F has. Thus it remains to verify that G is 2-l.s.c.

First we verify

CLAIM 1. *For each $\varepsilon>0$ and each $x \in X$,*

$$B_\varepsilon(G(x)) = B_\varepsilon(F(x)) \cap B_{\varepsilon+r}(f(x)).$$

For let $y \in B_\varepsilon(G(x))$. Then there exists $z \in F(x) \cap B_r(f(x))$ such that $y \in B_\varepsilon(z)$. Hence

$$y \in B_\varepsilon(F(x)) \cap B_\varepsilon(B_r(f(x))) = B_\varepsilon(F(x)) \cap B_{\varepsilon+r}(f(x))$$

and so $B_\varepsilon(G(x)) \subset B_\varepsilon(F(x)) \cap B_{\varepsilon+r}(f(x))$.

For the reverse inclusion, let $y \in B_\varepsilon(F(x)) \cap B_{r+\varepsilon}(f(x))$. Then $y = y_1 + e_1 = y_2 + e_2$, where $y_1 \in F(x)$, $y_2 \in B_r(f(x))$, and $e_i \in B_\varepsilon(0)$ for $i=1,2$. Since $F(x)$ and $B_r(f(x))$ are

both intervals in the 1-dimensional space Y and $F(x)\cap B_r(f(x))\neq\varnothing$, there exists λ, $0\leq\lambda\leq 1$, such that

$$\lambda y_1+(1-\lambda)y_2\in F(x)\cap B_r(f(x)).$$

Since $B_\varepsilon(0)$ is convex, $\lambda e_1+(1-\lambda)e_2\in B_\varepsilon(0)$ and so

$$\begin{aligned} y&=\lambda(y_1+e_1)+(1-\lambda)(y_2+e_2)=\lambda y_1+(1-\lambda)y_2+\lambda e_1+(1-\lambda)e_2\\ &\in F(x)\cap B_r(f(x))+B_\varepsilon(0)=B_\varepsilon(F(x)\cap B_r(f(x)))\\ &=B_\varepsilon(\overline{F(x)\cap B_r(f(x))})=B_\varepsilon(G(x)).\end{aligned}$$

This shows that $B_\varepsilon(F(x))\cap B_{\varepsilon+r}(f(x))\subset B_\varepsilon(G(x))$ and verifies the claim.

To prove that G is 2-l.s.c., let $x_0\in X$ and $\varepsilon>0$. Since F is 2-l.s.c. and f is continuous, there exists a neighborhood U of x_0 such that for all x_1, x_2 in U,

(i) $$B_\varepsilon(F(x_1))\cap B_\varepsilon(F(x_2))\neq\varnothing$$

and

(ii) $$f(x_1)\in B_\varepsilon(f(x_2)).$$

Using the claim, we have for all x_1, x_2 in U

$$B_\varepsilon(G(x_1))\cap B_\varepsilon(G(x_2))=B_\varepsilon(F(x_1))\cap B_{\varepsilon+r}(f(x_1))\cap B_\varepsilon(F(x_2))\cap B_{\varepsilon+r}(f(x_2)).$$

To show that this intersection of four sets is nonempty, it suffices by Helly's theorem [19, p. 196] to verify that any two of the sets has a nonempty intersection. Let $A_1=B_\varepsilon(F(x_1))$, $A_2=B_{\varepsilon+r}(f(x_1))$, $A_3=B_\varepsilon(F(x_2))$, and $A_4=B_{\varepsilon+r}(f(x_2))$. Since $B_\varepsilon(F(x))\cap B_{\varepsilon+r}(f(x))\supset F(x)\cap B_r(f(x))\neq\varnothing$, it follows that $A_1\cap A_2\neq\varnothing$ and $A_3\cap A_4\neq\varnothing$. By (i), $A_1\cap A_3\neq\varnothing$. Also (ii) implies that $A_1\cap A_4=B_\varepsilon(F(x_1))\cap B_{r+\varepsilon}(f(x_2))\supset F(x_1)\cap B_r(f(x_1))\neq\varnothing$. Similarly, $A_2\cap A_3\supset F(x_2)\cap B_r(f(x_2))\neq\varnothing$. Finally, $A_2\cap A_4\neq\varnothing$ by (ii). Since every two of the A_i's has a nonempty intersection, $\bigcap_1^4 A_i\neq\varnothing$. That is,

$$B_\varepsilon(G(x_1))\cap B_\varepsilon(G(x_2))\neq\varnothing$$

for all x_1, x_2 in U, and thus G is 2-l.s.c. at x_0. □

We can now state our first selection theorem, which shows that 2-lower semicontinuity is the essential property.

THEOREM 2.7. *Let X be a paracompact space, Y a 1-dimensional normed linear space, and suppose the mapping $F: X\to 2^Y$ has closed, bounded and convex images. Then F has a continuous selection if and only if F is 2-lower semicontinuous.*

Proof. Using Lemma 2.3, it suffices to verify sufficiency. Let F be 2-l.s.c. We will obtain a continuous selection for F as the limit of a certain sequence of functions. Towards this end, we will construct a sequence of continuous functions $f_k: X\to Y$ such that for all $x\in X$:

(i) $$f_k(x)\in B_{2^{-k}}(F(x)) \qquad (k=1,2,\cdots),$$

(ii) $$f_k(x)\in B_{4\cdot 2^{-k}}(f_{k-1}(x)) \qquad (k=2,3,\cdots).$$

We proceed by induction on k. By Theorem 2.5, F has a continuous 2^{-1}-approximate selection f_1. That is, $f_1: X\to Y$ is continuous and $f_1(x)\in B_{2^{-1}}(F(x))$ for all x. Next suppose that $f_1, f_2, \cdots, f_k$ have been chosen in accordance with (i) and (ii). Define

$$G(x)=\overline{F(x)\cap B_{2^{-k}}(f_k(x))}, \qquad x\in X.$$

By Lemma 2.6, G is 2-l.s.c. and has closed, bounded, and convex images. By Theorem 2.5 (applied to G instead of F), G has a continuous $2^{-(k+1)}$ approximate selection f_{k+1}.

In particular, $f_{k+1}: X \to Y$ is continuous and $f_{k+1}(x) \in B_{2^{-(k+1)}}(G(x))$ for all $x \in X$. Then for each $x \in X$,

$$f_{k+1}(x) \in B_{2^{-(k+1)}}(F(x))$$

and

$$f_{(k+1)}(x) \in B_{2^{-(k+1)}}(B_{2^{-k}}(f_k(x))) = B_{3 \cdot 2^{-(k+1)}}(f_k(x)) \subset B_{4 \cdot 2^{-(k+1)}}(f_k(x)).$$

This completes the induction step and hence establishes (i) and (ii).

By (ii), the sequence of functions $\{f_k\}$ is uniformly Cauchy so it must converge to some continuous function $f: X \to Y$. By (i), $f(x) \in F(x)$ for all $x \in X$. Thus f is a continuous selection for F. □

The idea of obtaining a continuous selection as a limit of continuous ε-approximate selections goes back at least to Michael [12]. Note that the example given at the end of this paper shows that lower semicontinuity *cannot* be substituted for 2-lower semicontinuity in Theorem 2.7.

In view of the improvement made in Theorem 2.5 in the case when $n=1$ (viz. Theorem 2.7) and noting the role played by Helly's theorem in the proof, it is natural to ask whether this improvement actually holds in the general case. That is, under the hypothesis of Theorem 2.5, if F is $(n+1)$-l.s.c., must F have a continuous selection? Pelant has kindly communicated a counterexample to us. His example is of a mapping F from $\mathbb{R}$ into the subsets of $\mathbb{R}^2$ which has closed, bounded and convex images, and has continuous ε-approximate selections for every $\varepsilon > 0$. But F has no continuous selection.

Before stating our next selection theorem which shows (again) that 2-lower semicontinuity is the essential property, we note the following useful equivalent reformulation of 2-lower semicontinuity.

LEMMA 2.8. *Let $F: X \to 2^Y$ and $x_0 \in X$. Consider the following statements:*

(1) *F is 2-lower semicontinuous at x_0.*

(2) *For each $\varepsilon > 0$, there is a neighborhood U of x_0 such that*

$$d(F(x_1), F(x_2)) \equiv \inf\{d(y_1, y_2) \mid y_i \in F(x_i),\ i = 1, 2\} < \varepsilon$$

for each choice of points x_1, x_2 in U.

(3) *For each pair of sequences $\{x_n\}$, $\{x_n'\}$ in X converging to x_0, there exist points $y_n \in F(x_n)$, $y_n' \in F(x_n')$ such that $d(y_n, y_n') \to 0$.*

Then (1) $\Leftrightarrow$ (2) $\Rightarrow$ (3). *If X is also a metric space, then* (3) $\Rightarrow$ (1) *and all three statements are equivalent.*

Remark. The property defined by statement (2) was first studied in [9] where it was called the "continuity property" (cp) of F at x_0. Thus (cp) and 2-lower semicontinuity are the same. Among other things, it was shown in [9] that many upper semicontinuous set-valued mappings must be (cp) on a dense G_δ set.

For a given mapping $F: X \to 2^Y$, let $S(F)$ denote the set of points for which F is single-valued; that is,

$$S(F) = \{x \in X \mid F(x) \text{ is a singleton}\}.$$

THEOREM 2.9. *Let Y be a complete metric space and suppose $F: X \to 2^Y$ has closed images and $S(F)$ is dense in X. Then F has a continuous selection if and only if F is 2-lower semicontinuous. Moreover, if F has a continuous selection, it is unique.*

Proof. The necessity has already been proved in Lemma 2.3. For the sufficiency, suppose F is 2-l.s.c. The proof will proceed through a series of claims.

CLAIM 2. *For each $x_0 \in X$ and each net $\{x_n\}$ in $S(F)$ with $x_n \to x_0$, the net $\{F(x_n)\}$ is Cauchy.*

For given any $\varepsilon>0$, choose a neighborhood U of x_0 such that $d(F(x),F(x'))<\varepsilon$ for all x,x' in U. Since $x_n\in U$ eventually, we have $d(F(x_n),F(x_m))<\varepsilon$ for all n and m eventually. Hence $\{F(x_n)\}$ is a Cauchy net.

CLAIM 3. *For each $x_0\in X$ and each net $\{x_n\}$ in $S(F)$ with $x_n\to x_0$, the limit $\lim F(x_n)$ exists, depends only on x_0 (and not the particular net converging to x_0), and $\lim F(x_n)\in F(x_0)$.*

Since Y is complete, Claim 2 implies that there exists $y_0\in Y$ such that $y_0=\lim F(x_n)$. If $\{z_m\}$ is any other net in $S(F)$ converging to x_0, then the same argument shows that there exists $y_0'\in Y$ such that $y_0'=\lim F(z_m)$. Given any $\varepsilon>0$, Lemma 2.8 implies that there exists a neighborhood U of x_0 such that $d(F(x),F(x'))<\varepsilon$ for all x,x' in U. Since x_n, z_m are in U eventually, $d(F(x_n),F(z_m))<\varepsilon$ eventually. Thus

$$d(y_0,y_0')=\lim_{n,m} d(F(x_n),F(z_m))\le\varepsilon.$$

Since ε was arbitrary, $y_0=y_0'$. Hence $\lim F(x_n)$ exists and depends only on x_0 and not the particular net in $S(F)$ converging to x_0. Finally, using Lemma 2.8, we get

$$d(y_0,F(x_0))=\lim d(F(x_n),F(x_0))=0.$$

Hence $y_0\in\overline{F(x_0)}=F(x_0)$.

Using Claim 3, we may define a function $f\colon X\to Y$ by

$$f(x)=\lim F(x_n),\qquad x\in X,$$

where $\{x_n\}$ is any net in $S(F)$ converging to x. Further, $f(x)\in F(x)$ for each $x\in X$. That is, f is a selection for F.

CLAIM 4. *For each $x_0\in X$,*

$$\inf_{U\in\mathfrak{U}(x_0)}\sup\{d(f(x_0),F(x))\,|\,x\in S(F)\cap U\}=0,$$

where $\mathfrak{U}(x_0)$ denotes the collection of all neighborhoods of x_0.

For if the result were false, there would exist a net $\{x_n\}$ in $S(F)$ with $x_n\to x_0$ and $d(f(x_0),F(x_n))\ge\varepsilon>0$ for all n. But this contradicts Claim 3.

CLAIM 5. *f is a continuous selection for F.*

Let $x_0\in X$ and $\varepsilon>0$ be given. By Claim 4, there exists a neighborhood U of x_0 such that $d(f(x_0),f(x))<\frac{\varepsilon}{2}$ for all $x\in S(F)\cap U$. Hence for any $x\in U$ (applying Claim 4 to x instead of x_0), there exists $x_1\in S(F)\cap U$ such that $d(f(x),f(x_1))<\frac{\varepsilon}{2}$. Thus

$$d(f(x_0),f(x))\le d(f(x_0),f(x_1))+d(f(x_1),f(x))<\varepsilon.$$

This proves f is continuous at x_0.

CLAIM 6. *f is the only continuous selection for F.*

This is a consequence of the fact that any other continuous selection for F must agree with f on the dense set $S(F)$, hence everywhere.

This completes the proof of the theorem. □

3. Applications to metric projections. Let M be a nonempty subset of the normed linear space X. For a given $x\in X$, the set of all *best approximations* or nearest points to x from M is defined by

$$P_M(x)=\{y\in M\,|\,\|x-y\|=d(x,M)\}.$$

M is called *proximinal* (resp. *Chebyshev*) provided $P_M(x)$ contains at least (resp. exactly) one point for each $x\in X$. For a proximinal set M, the set-valued mapping $P_M\colon X\to 2^M$ thus defined is called the *metric projection* onto M. It is well known and

easy to prove that a proximinal set M is always closed, $P_M(x)$ is always closed and bounded, and $P_M(x)$ is convex if M is convex.

As a consequence of Theorem 2.9, we have

COROLLARY 3.1. *Let M be a proximinal subset of the Banach space X and suppose that*

$$S(P_M)=\{x\in X \mid P_M(x) \text{ is a singleton}\}$$

is dense in X. Then P_M has a continuous selection if and only if P_M is 2-lower semicontinuous. Moreover, if P_M has a continuous selection, it is unique.

In the particular case that X is a strictly convex space, it is easy to verify that $S(P_M)$ is dense in X for any proximinal subset of M of X. (Indeed, if $x\in X\setminus M$ and $y\in P_M(x)$, then every point of the form $\lambda x+(1-\lambda)y$, $0<\lambda<1$, has y_0 as its *unique* best approximation in M. In particular, there are points arbitrarily close to x lying in $S(P_M)$.) If, moreover, P_M admits a continuous selection, then M must be a Chebyshev set, i.e. $S(P_M)=X$. (For if $x\in X\setminus S(P_M)$, choose distinct points y_1, y_2 in $P_M(x)$. Then the sequences $x_n=(1-\frac{1}{n})x+\frac{1}{n}y_1$ and $z_n=(1-\frac{1}{n})x+\frac{1}{n}y_2$ have the property that $P_M(x_n)=y_1$ and $P_M(z_n)=y_2$ for n sufficiently large, and both $\{x_n\}$ and $\{z_n\}$ converge to x. This precludes the existence of a selection for P_M which is continuous at x.) These remarks prove that *if M is a proximinal subset of the strictly convex space X, then P_M has a continuous selection if and only if M is Chebyshev and P_M is continuous.*

A subset M of the normed linear space X is called *almost Chebyshev* if $X\setminus S(P_M)$ is a set of first category in X. Garkavi [7], [8] studied almost Chebyshev sets and showed that they are rather plentiful. For example, in every separable Banach space (resp. dual space), there exist almost Chebyshev subspaces of every finite dimension (resp. finite codimension). Since the complement of a first category set in a Banach space is dense, Theorem 2.9 implies the following result.

COROLLARY 3.2. *Let M be a proximinal almost Chebyshev subset of the Banach space X. Then P_M has a continuous selection if and only if P_M is 2-lower semicontinuous. Moreover, if P_M has a continuous selection, it is unique.*

Let T be a compact Hausdorff space. A finite dimensional subspace M of $C(T)$ is called a *Z-subspace* [11] if the only element of M which vanishes on an open subset of T is the zero function. As a consequence of the main result of Garkavi [8], Z-subspaces are almost Chebyshev. This fact and Corollary 3.2 yield:

COROLLARY 3.3. *Let M be a Z-subspace of $C(T)$. Then P_M has a continuous selection if and only if P_M is 2-lower semicontinuous. If P_M has a continuous selection, it is unique.*

The uniqueness statement in this corollary was also proved by Brown [2] without appealing to Garkavi's theorem.

In related work, Brown [3] proved a "Mairhuber-type" theorem when he showed that if $C(T)$ contains a Z-subspace M of dimension at least two and P_M admits a continuous selection, then T is essentially an "interval." Nürnberger [14], building on earlier work in [16] and [22], gave an intrinsic characterization of those Z-subspaces of $C[a,b]$ whose metric projections admit continuous selections. More generally, Sommer [21] filled the remaining gap left by [14], [16], and [22] and thus completed an intrinsic characterization of those finite dimensional subspaces of $C[a,b]$ whose metric projections admit continuous selections. The more general question [11] of characterizing those n (>1)-dimensional subspaces of $C(T)$ whose metric projections admit continuous selections remains open.

Our next result is an immediate consequence of Theorems 2.4 and 2.5.

COROLLARY 3.4. *Let M be an n-dimensional subspace of the normed linear space X. Then the following statements are equivalent.*

(1) *P_M has a continuous ε-approximate selection for each $\varepsilon>0$.*

(2) *P_M is almost lower semicontinuous.*

(3) *P_M is $(n+1)$-lower semicontinuous.*

We have also the following consequence of Theorem 2.7.

COROLLARY 3.5. *Let M be a 1-dimensional subspace of the normed linear space X. Then P_M has a continuous selection if and only if P_M is 2-lower semicontinuous.*

In relation to this corollary, we note that Lazar, Morris and Wulbert [11] gave an intrinsic characterization of those 1-dimensional subspaces of $C(T)$ whose metric projections admit continuous selections. Also, Lazar [10] gave an intrinsic characterization of those 1-dimensional subspaces of l_1 whose metric projections admit continuous selections.

We conclude the paper with an example which sheds some further light on the distinction between 2-lower semicontinuity and lower semicontinuity, and shows that some of the more obvious choices for continuous selections fail.

Example. Let $X=\mathbb{R}^3$ with the norm

$$\|x\|=|x(1)|+\sqrt{x(2)^2+x(3)^2}, \qquad x=(x(1),x(2),x(3)).$$

(The unit ball looks like a "double cone.") Let M be the 1-dimensional subspace $M=\operatorname{span}\{x_1\}$, where $x_1=(1,1,0)$. It is easy to verify that

$$P_M(x)=\begin{cases} x(1)x_1 & \text{if } x(3)\neq 0, \\ \{\alpha x_1 \mid \alpha=\lambda x(1)+(1-\lambda)x(2), 0\leq\lambda\leq 1\} & \text{if } x(3)=0. \end{cases}$$

In particular, the function

$$f(x)=x(1)x_1, \qquad x\in X$$

is a continuous (even linear!) selection for P_M. Since $S(P_M)=\{x\in X \mid x(3)\neq 0\}$ is dense in X, Corollary 3.1 implies that P_M is 2-l.s.c. and f is the only continuous selection for P_M. Observe, however, that P_M is *not* lower semicontinuous at the point $x_0=(1,0,0)$. (For the open set $V=\{x\in X \mid \|x\|<1\}$ has the property that $P_M(x_0)\cap V_n\neq\varnothing$ since 0 is in the intersection. Yet for each neighborhood U of x_0, the element $z_n=(1,n^{-1},n^{-1})$ is in U eventually and $P_M(z_n)\cap V=\{x_1\}\cap V=\varnothing$.)

Furthermore, f is not obtained by any of the more natural choices for a selection. That is, in general, $f(x)$ is *not* the element of $P_M(x)$ having minimal norm (for $\|f(x_0)\|=1>0=\inf\{\|y\| \mid y\in P_M(x_0)\}$); $f(x)$ is *not* the midpoint of the line segment $P_M(x)$ (for the midpoint of $P_M(x_0)$ is $\frac{1}{2}x_1\neq x_1=f(x_0)$); $f(x)$ is *not* the element of $P_M(x)$ with the largest (or smallest) coefficient of x_1 (for the element of $P_M(-x_0)$ which has the largest coefficient is $0\neq -x_1=f(-x_0)$, and the element of $P_M(x_0)$ which has the smallest coefficient is $0\neq x_1=f(x_0)$).

Acknowledgment. Both authors are particularly grateful to Professor Bruno Brosowski for his kind assistance and generosity during our stay in Frankfurt.

REFERENCES

[1] A. L. BROWN, *Best n-dimensional approximation to sets of functions*, Proc. London Math. Soc., 14 (1964), pp. 577–594.

[2] _____, *On continuous selections for metric projections in spaces of continuous functions*, J. Functional Anal., 8 (1971), pp. 431–449.

[3] _____, *An extension to Mairhuber's theorem: On metric projections and discontinuity of multivariate best uniform approximation*, J. Approx. Theory, to appear.

[4] A. CELLINA, *A theorem on the approximation of compact multi-valued mappings*, Accad. Naz. Lincei Rend. Cl. Sc. Fis. Mat. Natur., 47 (1969), pp. 429–433.

[5] _____, *A further result on the approximation of set-valued mappings*, Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 48 (1970), pp. 230–234.

[6] F. DEUTSCH AND P. KENDEROV, *When does the metric projection admit a continuous selection?*, in Approximation Theory III, E. W. Cheney, ed., Academic Press, New York, 1980, pp. 327–333.

[7] A. L. GARKAVI, *On Čebyšev and almost Čebyšev subspaces*, Izv. Akad. Nauk SSSR Ser. Mat., 28 (1964), pp. 799–818. (Translation in Amer. Math. Soc. Transl., 96 (1970), pp. 153–175.)

[8] _____, *Almost Čebyšev systems of continuous functions*, Izv. Vyss. Učebn. Zaved. Matematika, 45 (1965), pp. 36–44. (Translation in Amer. Math. Soc. Transl., 96 (1970), pp. 177–187.)

[9] P. KENDEROV, *Dense strong continuity of pointwise continuous mappings*, Pacific J. Math., 89 (1980), pp. 111–130.

[10] A. J. LAZAR, *Spaces of affine continuous functions on simplexes*, Trans. Amer. Math. Soc., 134 (1968), pp. 503–525.

[11] A. J. LAZAR, P. D. MORRIS AND D. E. WULBERT, *Continuous selections for metric projections*, J. Functional Anal., 3 (1969), pp. 193–216.

[12] E. MICHAEL, *Selected selection theorems*, Amer. Math. Monthly, 63 (1956), pp. 233–237.

[13] G. NÜRNBERGER, *Continuous selections for the metric projection and alternation*, J. Approx. Theory, 28 (1980), pp. 212–226.

[14] _____, *Nonexistence of continuous selections for the metric projection and weak Chebyshev subspaces*, this Journal, 11 (1980), pp. 460–467.

[15] _____, *Schnitte für die metrische Projektion*, J. Approx. Theory, 20 (1977), pp. 196–219.

[16] G. NÜRNBERGER AND M. SOMMER, *Weak Chebyshev subspaces and continuous selections for the metric projection*, Trans. Amer. Math. Soc., 238 (1978), pp. 129–138.

[17] _____, *Characterization of continuous selections of the metric projection for spline functions*, J. Approx. Theory, 22 (1978), pp. 320–330.

[18] S. REICH, *Approximate selections, best approximations, fixed points, and invariant sets*, J. Math. Anal. Appl., 62 (1978), pp. 104–113.

[19] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[20] M. SOMMER, *Characterization of continuous selections for the metric projection for generalized splines*, this Journal, 11 (1980), pp. 23–40.

[21] _____, *Continuous selections for metric projection*, in Quantitative Approximation, Bonn 1979, R. De Vore and K. Scherer, eds., Academic Press, New York, 1980, pp. 301–317.

[22] _____, *Nonexistence of continuous selections of the metric projection for a class of weak Chebyshev spaces*, Trans. Amer. Math. Soc., 260 (1980), pp. 403–409.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0016 $01.25/0

A COMPARISON OF SOME MULTIVARIATE PADÉ-APPROXIMANTS*

ANNIE A. M. CUYT†

Abstract. In [5] Levin defined general order Padé-type rational approximants of a function of n variables (here referred to as "type L" approximants). In §1 we repeat briefly the defining equations and the determinant representation for $n=2$. Levin proved that the Chisholm approximants were a special case of his "type L" approximants.

The multivariate Padé-approximants (here referred to as "type C" approximants) were introduced in [1] and [2]; we repeat the definition for $n=2$ in §2. They have several nice properties which often imply numerical advantages; examples of such situations are given in [3] and [4].

In §3 we show that "type C" approximants are also a special case of the "type L" approximants. The explicit determinant formulas are a link between the solution of the Padé approximation problem and the irreducible rational form of the solution. Via the determinant representation we can also see that, in the case of "type C" approximants, we deal with matrices that are near-Toeplitz. This is not true for the Chisholm approximants. A theorem concerning the displacement-rank of the matrix of the homogeneous system, defining the coefficients of the denominator of the "type C" approximant, is proved.

In §4 analogous results are formulated for $n>2$.

1. General order Padé-type rational approximants in two variables (or type L approximants). We repeat some notations and definitions given by Levin.

Let $\mathbf{N}=\{0,1,2,\cdots\}$. Given a subset D of $\mathbf{Z}^2$ we define:

(a) the complement $\bar{D}=\mathbf{Z}^2\setminus D$,

(b) the (i,j)-translation of D as $D_{ij}=\{(k,n)\,|\,(k+i,n+j)\in D\}$,

(c) the nonnegative part of D as $D^+=D\cap\mathbf{N}^2$.

To any subset D such that D^+ is a finite set we associate polynomials

$$\sum_{(i,j)\in D^+} b_{ij}x^iy^j \quad \text{with } b_{ij} \text{ in } \mathbf{R}$$

We call D the *rank* of the polynomials. Given the double power series

$$f(x,y)=\sum_{i,j=0}^{\infty} c_{ij}x^iy^j$$

we will choose three subsets N, D and E of $\mathbf{Z}^2$ and construct an $[N/D]_E$ approximation to $f(x,y)$ as follows:

$$P(x,y)=\sum_{(i,j)\in N^+} a_{ij}x^iy^j \qquad (N\leftarrow\text{numerator}), \tag{1.1a}$$

$$Q(x,y)=\sum_{(i,j)\in D^+} b_{ij}x^iy^j \qquad (D\leftarrow\text{denominator}), \tag{1.1b}$$

$$(f\cdot Q-P)(x,y)=\sum_{(i,j)\in \bar{E}^+} d_{ij}x^iy^j \qquad (E\leftarrow\text{equations}). \tag{1.1c}$$

We select N, D and E such that

(a) $D\subset\mathbf{N}^2$ has m elements, numbered

$$(i_1,j_1),\cdots,(i_m,j_m),$$

*Received by the editors April 28, 1981, and in revised form October 27, 1981.

†Department of Mathematics, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium. Aspirant N.F.W.O.

(b) $N \subset E$ and $H = E \backslash N$ has $m-1$ elements in $\mathbf{N}^2$, numbered

$$(h_2,k_2),\cdots,(h_m,k_m) \qquad (H \leftarrow \text{homogeneous equations}).$$

Then $P(x,y)$ and $Q(x,y)$, defined by (1.1c), are given by:

$$P(x,y)=\begin{vmatrix} x^{i_1}y^{j_1}N_{i_1j_1}(x,y) & x^{i_2}y^{j_2}N_{i_2j_2}(x,y) & \cdots & x^{i_m}y^{j_m}N_{i_mj_m}(x,y) \\ c_{h_2-i_1,k_2-j_1} & c_{h_2-i_2,k_2-j_2} & \cdots & c_{h_2-i_m,k_2-j_m} \\ c_{h_3-i_1,k_3-j_1} & c_{h_3-i_2,k_3-j_2} & \cdots & c_{h_3-i_m,k_3-j_m} \\ \vdots & \vdots & & \vdots \\ c_{h_m-i_1,k_m-j_1} & c_{h_m-i_2,k_m-j_2} & \cdots & c_{h_m-i_m,k_m-j_m} \end{vmatrix}$$

where

$$N_{i_1j_1}(x,y)=\sum_{(i,j)\in N^+_{i_1j_1}} c_{ij}x^iy^j,$$

and

$$Q(x,y)=\begin{vmatrix} x^{i_1}y^{j_1} & x^{i_2}y^{j_2} & \cdots & x^{i_m}y^{j_m} \\ c_{h_2-i_1,k_2-j_1} & c_{h_2-i_2,k_2-j_2} & \cdots & c_{h_2-i_m,k_2-j_m} \\ c_{h_3-i_1,k_3-j_1} & c_{h_3-i_2,k_3-j_2} & \cdots & c_{h_3-i_m,k_3-j_m} \\ \vdots & \vdots & & \vdots \\ c_{h_m-i_1,k_m-j_1} & c_{h_m-i_2,k_m-j_2} & \cdots & c_{h_m-i_m,k_m-j_m} \end{vmatrix}. \tag{1.1d}$$

2. Multivariate Padé-approximants for a double power series (or type C approximants). We define a polynomial of degree l in two variables as

$$\sum_{i+j=0}^{l} a_{ij}x^iy^j.$$

A term $a_{ij}x^iy^j$ is said to be of degree $i+j$. The order $\partial_0 P$ and the exact degree ∂P are defined by

$$\partial_0 P=\min\{i+j \mid a_{ij}\neq 0\}, \qquad \partial P=\max\{i+j \mid a_{ij}\neq 0\}.$$

In the Padé-approximation problem of order (l,m) we try to find a pair (P,Q) of two-variable polynomials,

$$P(x,y)=\sum_{i+j=lm}^{lm+l} a_{ij}x^iy^j, \tag{2.1a}$$

$$Q(x,y)=\sum_{i+j=lm}^{lm+m} b_{ij}x^iy^j, \tag{2.1b}$$

such that

$$(f\cdot Q-P)(x,y)=\sum_{i+j=lm+l+m+1}^{\infty} d_{ij}x^iy^j. \tag{2.1c}$$

Equation (2.1c) is equivalent with

$$\partial_0(f\cdot Q-P)\geq lm+l+m+1.$$

A nontrivial $Q(x,y)$, such that (2.1) is satisfied, always exists [2]. If the polynomials $R(x,y)$ and $S(x,y)$ also satisfy (2.1), in other words if $\partial_0(f\cdot S-R)\geq lm+l+m+1$, too, then

$$P(x,y)\cdot S(x,y)=Q(x,y)\cdot R(x,y).$$

This property justifies the following definitions:

(a) Let $(P_*/Q_*)(x,y)$ be the irreducible form of $(P/Q)(x,y)$ such that $Q_*(0,0)=1$; if this form exists we call it the multivariate Padé-approximant of order (l,m) for f.

(b) If the irreducible form $(P_*/Q_*)(x,y)$ is such that $\partial_0 Q_*\geq 1$, then we call P_*/Q_* the multivariate rational approximant of order (l,m) for f.

The (l,m)-multivariate rational approximant is unique up to a multiplicative constant in numerator and denominator. For $P_*(x,y)$ and $Q_*(x,y)$ we define:

$$l'=\partial P_*-\partial_0 Q_*,\qquad m'=\partial Q_*-\partial_0 Q_*.$$

We can prove that

$$l'\leq l,\qquad m'\leq m.$$

It is also easy to verify the following theorem.

THEOREM 2.1. *For the irreducible form P_*/Q_* of P/Q where (P,Q) satisfies* (2.1) *and for every polynomial $R(x,y)=\sum_{i=0}^{s} r_i x^i y^{s-i}$ with $s=lm-\partial_0 Q_*+\min(l-l',m-m')$, $(P_*\cdot R, Q_*\cdot R)$ satisfies* (2.1).

Also $s=lm-\partial_0 Q_*+\min(l-l',m-m')$ is the highest possible degree that allows the construction of a homogeneous polynomial $R(x,y)=\sum_{i=0}^{s} r_i x^i y^{s-i}$ such that (2.1) is satisfied by $(P_*\cdot R, Q_*)\cdot R$. From now on the multivariate Padé-approximant as well as the multivariate rational approximant will be called type C approximants.

3. Connection between the two approaches. First of all we remark that for the case of one variable the type-L approximant [5] as well as the type C approximant [1,2] reduce to the well-known ordinary Padé-approximant. And the polynomials $P(x,y)$ and $Q(x,y)$ satisfying (2.1) do also satisfy (1.1) when the sets N, D and E are chosen as follows:

$$N=\{(i,j)\,|\,i,j\in\mathbf{N},\ lm\leq i+j\leq lm+l\},$$
$$D=\{(i,j)\,|\,i,j\in\mathbf{N},\ lm\leq i+j\leq lm+m\},$$
$$E=\{(i,j)\,|\,i,j\in\mathbf{N},\ lm\leq i+j\leq lm+l+m\}.$$

The set $H=E\backslash N$ has one element less than the set D, as required; but we could also add to E the set $\{(i,j)\,|\,i,j\in\mathbf{N},\ i+j<lm\}$, since $\partial_0(f\cdot Q-P)\geq lm$ for all polynomials P and Q as in (2.1a) and (2.1b). Doing so we do not impose more conditions on the coefficients a_{ij} and b_{ij}; we write

$$E^{\text{ext}}=\{(i,j)\,|\,i,j\in\mathbf{N},\ i+j\leq lm+l+m\}.$$

Let us now number the points in D and H, using a diagonal enumeration:

(a)

$$D=\Big\{\underbrace{(lm,0),(lm-1,1),\cdots,(0,lm)}_{\text{first diagonal}},\ \underbrace{(lm+1,0),\cdots,(0,lm+1)}_{\text{second diagonal}},\cdots,$$
$$\underbrace{(lm+m,0),\cdots,(0,lm+m)}_{\text{last diagonal}}\Big\},$$

(b)

$$H=\Big\{\underbrace{(lm+l+1,0),(lm+l,1),\cdots,(0,lm+l+1)}_{\text{first diagonal}},\underbrace{(lm+l+2,0),\cdots,(0,lm+l+2)}_{\text{second diagonal}},\cdots,$$

$$\underbrace{(lm+l+m,0),\cdots,(0,lm+l+m)}_{\text{last diagonal}}\Big\}.$$

When we write down the equations equivalent with condition (2.1c), the set of homogeneous equations in the unknown b_{ij} has a coefficient matrix which is exactly the matrix in (1.1d) after removing the first row. From now on we will call this matrix $\mathcal{H}$; it has $p=\binom{lm+l+m+2}{2}-\binom{lm+l+2}{2}$ rows and one more columns than rows.

THEOREM 3.1. *The rank of the matrix* $\mathcal{H}$ *is at most* $p-(lm-\partial_0 Q_*+\min(l-l', m-m'))$.

Proof. We only have to prove that the dimension of the null-space of $\mathcal{H}$, which is the dimension of the space of solutions of the homogeneous system of equations, is at least $lm-\partial_0 Q_*+\min(l-l',m-m')+1$; in other words that (2.1) admits solutions where at least $lm-\partial_0 Q_*+\min(l-l',m-m')+1$ of the b_{ij} can be freely chosen. Precisely this is formulated in Theorem 2.1.

The number $s=lm-\partial_0 Q_*+\min(l-l',m-m')$ is one less than the number of coefficients in a homogeneous polynomial of degree s in two variables, namely $\binom{s+1}{s}$. The number of coefficients in a homogeneous polynomial of degree s in n variables is $\binom{s+n-1}{s}$. But first of all we are going to take a closer look at the matrix $\mathcal{H}$ for the type C approximants when $n=2$; in the next section we will treat the n-variable case with $n>2$. To examine the special structure of $\mathcal{H}$ we introduce the following notation. For $Q(x,y)=\sum_{i+j=lm}^{lm+m} b_{ij}x^i y^j$ we write

$$B_{lm}=\begin{pmatrix} b_{lm,0} \\ b_{lm-1,1} \\ \vdots \\ b_{0,lm} \end{pmatrix},\quad B_{lm+1}=\begin{pmatrix} b_{lm+1,0} \\ b_{lm,1} \\ \vdots \\ b_{0,lm+1} \end{pmatrix},\quad \cdots,\quad B_{lm+m}=\begin{pmatrix} b_{lm+m,0} \\ b_{lm+m-1,1} \\ \vdots \\ b_{0,lm+m} \end{pmatrix}.$$

Equations (2.1c) can now be written as

$$\begin{pmatrix} H_{l+1,lm} & H_{l,lm+1} & \cdots & H_{l+1-m,lm+m} \\ H_{l+2,lm} & \cdots & & \\ \vdots & & & \\ H_{l+m,lm} & \cdots & & H_{l,lm+m} \end{pmatrix}\begin{pmatrix} B_{lm} \\ \vdots \\ B_{lm+m} \end{pmatrix}=0,$$

where $H_{i,j}$ is a matrix with $(i+j+1)$ rows and $(j+1)$ columns and the first column equal to the transpose of $(c_{i,0}c_{i-1,1}\cdots c_{1,i-1}c_{0,i}0\cdots 0)$ and the next columns equal to their previous column but with all the elements shifted down one place and a zero added on top.

To calculate the displacement rank $\alpha(\mathcal{H})$ of $\mathcal{H}$, we have to construct the lower shifted difference

$$\mathcal{H}-\overline{\mathcal{H}}=\begin{pmatrix} h_{1,1} & \cdots & h_{1,p+1} \\ \vdots & & \\ h_{p,1} & & h_{p,p+1} \end{pmatrix}-\begin{pmatrix} 0 & \cdots & & 0 \\ & h_{1,1} & \cdots & h_{1,p} \\ \vdots & & & \\ & \vdots & & \\ 0 & h_{p-1,1} & \cdots & h_{p-1,p} \end{pmatrix}$$

$$=\begin{pmatrix} h_{1,1} & \cdots & h_{1,p+1} \\ \vdots & & \\ & & \delta\mathcal{H} \\ h_{p,1} & & \end{pmatrix}.$$

Now $\alpha(\mathcal{H})=\operatorname{rank}(\delta\mathcal{H})+2$ [6].

THEOREM 3.2. *The displacement-rank of the matrix $\mathcal{H}$ is at most $m+2$.*

Proof. Let us write down the matrix $\mathcal{H}$ more explicitly:

$$\mathcal{H}=\left(\begin{array}{cccc|cccc|c|cccc}
c_{l+1,0} & 0 & \cdots & 0 & c_{l,0} & 0 & \cdots & 0 & & c_{l+1-m,0} & 0 & \cdots & 0 \\
\vdots & & & \vdots & \vdots & & & \vdots & & \vdots & & & \vdots \\
 & \ddots & & 0 & & \ddots & & 0 & & & \ddots & & 0 \\
c_{0,l+1} & & & c_{l+1,0} & c_{0,l} & & & c_{l,0} & & c_{0,l+1-m} & & & c_{l+1-m,0} \\
0 & & & \vdots & 0 & & & \vdots & \cdots & 0 & & & \vdots \\
\vdots & \ddots & & & \vdots & \ddots & & & & \vdots & \ddots & & \\
0 & \cdots & 0 & c_{0,l+1} & 0 & \cdots & 0 & c_{0,l} & & 0 & \cdots & 0 & c_{0,l+1-m} \\
\hline
 & & \vdots & & & & & & & & & & \\
\hline
c_{l+m,0} & 0 & \cdots & 0 & & & & & & & & & \\
\vdots & & & \vdots & & & & & & & & & \\
 & \ddots & & 0 & & & & & & & & & \\
c_{0,l+m} & & & c_{l+m,0} & & & & & & & & & \\
0 & & & \vdots & & & & & & & & & \\
\vdots & \ddots & & & & & & & & & & & \\
0 & \cdots & 0 & c_{0,l+m} & & & & & & & & &
\end{array}\right)$$

In our case $\delta\mathcal{H}$ has the structure $\delta\mathcal{H}=(\Delta_1\Delta_2\cdots\Delta_{m+1})$, where Δ_1 has $(p-1)$ rows and lm columns, Δ_i has $(p-1)$ rows and $(lm+i)$ columns for $i=2,\cdots,m+1$, and only the first column in Δ_i with $i\geq 2$ may contain nonzero elements; all the other elements in $\delta\mathcal{H}$ equal zero. So $\operatorname{rank}(\delta\mathcal{H})\leq m$ and this proves our theorem.

We will illustrate the theorems with some very simple examples. Consider $f(x,y)=1+x/(0\cdot 1-y)+\sin(xy)$.

(a) Take $l=1=m$. The type C approximant is

$$\frac{1+10x-10\cdot 1y}{1-10\cdot 1y}$$

with $l'=1=m'$, $\partial_0 Q_*=0$, $s=1$ and $\alpha(\mathcal{H})=3$.

The matrix

$$\mathcal{H}=\begin{bmatrix} 0 & 0 & 10 & 0 & 0 \\ 101 & 0 & 0 & 10 & 0 \\ 0 & 101 & 0 & 0 & 10 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix};$$

its rank is 3.

(b) Take $l=4$ and $m=2$. The type C approximant is

$$\frac{1+10x-10y+xy-10xy^2}{1-10y}$$

with $l'=3$, $m'=1$, $\partial_0 Q_*=0$, $s=9$ and $\alpha(\mathcal{H})=4$.

The matrix

$$\mathcal{H}=\begin{bmatrix} H_1 & H_2 & H_3 \\ H_4 & H_5 & H_6 \end{bmatrix},$$

where

$$H_1=10^5[\delta_{i,j+4}], \text{ a } 14\times 9 \text{ matrix},$$
$$H_2=10^4[\delta_{i,j+3}], \text{ a } 14\times 10 \text{ matrix},$$
$$H_3=10^3[\delta_{i,j+2}], \text{ a } 14\times 11 \text{ matrix},$$
$$H_4=10^6[\delta_{i,j+5}]-\tfrac{1}{6}[\delta_{i,j+3}], \text{ a } 15\times 9 \text{ matrix},$$
$$H_5=10^5[\delta_{i,j+4}], \text{ a } 15\times 10 \text{ matrix},$$
$$H_6=10^4[\delta_{i,j+3}], \text{ a } 15\times 11 \text{ matrix},$$

and $\delta_{i,j}$ is the Kronecker symbol (here used in rectangular matrices). $\mathcal{H}$ is a matrix of rank 20.

(c) Take $l=1$ and $m=2$. The type C approximant is

$$\frac{x-1.01y+10y^2+10x^2-20.2xy}{x-1.01y+10y^2-10.1xy+2.01xy^2}$$

with $l'=1$, $m'=2$, $\partial_0 Q_*=1$, $s=1$ and $\alpha(\mathcal{H})=4$. The matrix

$$\mathcal{H}=\left[\begin{array}{ccc:ccc:ccccc} 0 & 0 & 0 & & 10I_4 & & & & I_5 & & \\ & 101I_3 & & & & & & & & & \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 10I_5 & & \\ \hdashline 0 & 0 & 0 & & 10I_4 & & & & & & \\ & 1000I_3 & & & & & & & & & \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

where I_k is the $k\times k$ unit-matrix. The rank of $\mathcal{H}$ is 10.

(d) Consider $f(x,y)=(xe^x-ye^y)/x-y$ and take $l=1=m$. The determinant representations yield

$$P(x,y)=-\tfrac{1}{2}(x+y+0.5x^2+1.5xy+0.5y^2),$$
$$Q(x,y)=-\tfrac{1}{2}(x+y-0.5x^2-0.5xy-0.5y^2),$$

and indeed the type C approximant is

$$\frac{P(x,y)}{Q(x,y)}=\frac{P_*(x,y)}{Q_*(x,y)}$$

with $\partial_0 Q_*=1$ and $l'=1=m'$. The matrix $\mathcal{H}$ has rank p ($s=0$) and $\alpha(\mathcal{H})=3$.

4. The multivariate case. Given the power series

$$f(x)=\sum_{k=0}^{\infty} c_k x^k,$$

where $x=(x_1,\cdots,x_n)\in\mathbf{R}^n$, $c_k=c_{k_1\cdots k_n}$, $x^k=x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n}$ and $\Sigma_{k=0}^{\infty}=\Sigma_{k_1=0}^{\infty}\cdots\Sigma_{k_n=0}^{\infty}$, the Padé-approximation problem of order (l,m) is the following:

find

$$P(x)=\sum_{|i|=lm}^{lm+l} a_i x^i, \tag{4.1a}$$

$$Q(x)=\sum_{|j|=lm}^{lm+m} b_j x^j \tag{4.1b}$$

where $a_i=a_{i_1\cdots i_n}$ and $b_j=b_{j_1\cdots j_n}$, $|i|=i_1+\cdots+i_n$ and $|j|=j_1+\cdots+j_n$, such that

$$\partial_0(f\cdot Q-P)\geq lm+l+m+1 \tag{4.1c}$$

where ∂_0 is again the degree of the first nonzero term.

After calculation of the nontrivial solution of (4.1) [2] we can proceed as in §2 and define the multivariate Padé-approximant of order (l,m) and the multivariate rational approximant of order (l,m).

The integers l' and m' are defined as in the two-variable case and it is easy to prove the following n-dimensional analogue of Theorem 2.1.

THEOREM 4.1. *For the irreducible form P_*/Q_* of P/Q where (P,Q) satisfies* (4.1) *and for every polynomial $R(x)=\Sigma_{|i|=s}r_i x^i$ with $s=lm-\partial_0 Q_*+\min(l-l',m-m')$, $(P_*\cdot R,Q_*\cdot R)$ satisfies* (4.1).

Let us again study the connection with the approach of Levin. Condition (4.1c) results in $\binom{n+lm+l+m}{lm+\ l+m}$ equations: the first $\binom{n+lm+l}{lm\ +l}$ equations express the a_i as linear combinations of the b_j and the remaining equations form an overdetermined homogeneous linear system in the unknown b_j [2]; there are

$$p+1=\binom{n+lm+m}{lm+m}-\binom{n+lm-1}{lm-1}$$

unknown coefficients b_j. The b_j can be found by solving a homogeneous subsystem of p equations, having the rank of the overdetermined system.

Choose the sets N,D and H as follows:

- $N=\{i=(i_1,\cdots,i_n)\,|\,i\in\mathbf{N}^n,\ lm\leq|i|\leq lm+l\}$;
- $D=\{i=(i_1,\cdots,i_n)\,|\,i\in\mathbf{N}^n,\ lm\leq|i|\leq lm+m\}$;
- select a particular b_j and let $c_{h(k)-j}$ be the coefficient of b_j in the kth equation of the homogeneous subsystem we have to solve ($k=1,\cdots,p$),

$$H=\{h(k)=(h_1(k),\cdots,h_n(k))\,|\,k=1,\cdots,p\}.$$

We call the coefficient matrix of the homogeneous subsystem again $\mathcal{H}$. It is easy to prove the following n-dimensional analogue of Theorem 3.1.

THEOREM 4.2. *The rank of the matrix* $\mathcal{H}$ *is at most* $p-\binom{s+n-1}{s}+1$ *with* $s=lm-\partial_0 Q_*+\min(l-l', m-m')$.

If we use an enumeration of the points in D and H, similar to the one described in §3, it is obvious that in the multivariate case $\mathcal{H}$ is also a matrix with low displacement rank.

Acknowledgment. I want to express my sincere thanks to the referee for constructive remarks.

REFERENCES

[1] ANNIE A. M. CUYT, *Abstract Padé-approximants in operator theory*, Lecture Notes in Mathematics, 765 L. Wuytack ed., Springer, Berlin, 1979, pp. 61–87.

[2] _____, *Multivariate Padé-approximants*, J. Math. Anal. Appl., to appear.

[3] _____, *Numerical comparison of abstract Padé-approximants and abstract rational approximants with other generalizations of the classical Padé-approximant*, Lecture Notes in Mathematics 888, H. van Rossum and M. de Bruin, eds., Springer, Berlin, 1981, pp. 137–157.

[4] ANNIE A. M. CUYT AND P. VAN DER CRUYSSEN, *Abstract Padé-approximants for the solution of a system of nonlinear equations*, Computers and Mathematics with Applications, to appear.

[5] D. LEVIN, *General order Padé-type rational approximants defined from double power series*, J. Inst. Math. Appl., 18 (1976), pp. 1–8.

[6] M. MORF, S. KUNG AND T. KAILATH, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

SIAM J. MATH. ANAL.
Vol. 14, No. 1, January 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1401-0017 $01.25/0

TYPE t ENTROPY AND MAJORIZATION*

A. CLAUSING†

Abstract. Let $t>0$ and let $\mathbf{p}=(p_1,\cdots,p_n)$ be a probability vector. By $H_t(\mathbf{p})=-(\sum_{i=1}^n p_i^t \log p_i)/(\sum_{i=1}^n p_i^t)$ we denote the type t entropy of $\mathbf{p}$. Recently, Stolarsky has considered the problem of finding the best, that is, smallest, value of t such that the entropy inequality $H_t(p)\leq \log n$ is valid for all $\mathbf{p}$. In this paper we determine this value. We also prove that for $n\geq 3$ and optimal t equality can be achieved in the above inequality with a probability vector different from the trivial one $\mathbf{p}=(\frac{1}{n},\cdots,\frac{1}{n})$. This confirms a conjecture of Stolarsky.

1. Introduction. Let $\mathbf{p}=(p_1,\cdots,p_n)$ be a probability vector, that is, $p_i\geq 0$ $(i=1,\cdots,n)$ and $\sum_{i=1}^n p_i=1$. Also, let $t>0$. The quantity

$$H_t(\mathbf{p})=-\frac{\sum_{i=1}^n p_i^t \log p_i}{\sum_{i=1}^n p_i^t} \tag{1}$$

is known as Kapur's entropy of order 1 and type t. (If $p_i=0$, put $p_i^t \log p_i=0$ in the sum.) It has been discussed by various authors (cf. [1], [3], [4]). In particular, a result of Kapur ([1, p. 433, let α tend to 1 in line (19)]) implies that H_t achieves its maximum value at $\mathbf{p}=(\frac{1}{n},\cdots,\frac{1}{n})$, that is,

$$H_t(\mathbf{p})\leq \log n \quad \text{for all } \mathbf{p}. \tag{2}$$

For $t=1$, this is the well-known entropy inequality due to Shannon. Since $H_t(\mathbf{p})$ is decreasing in t, (2) is a stronger result than Shannon's inequality if $t<1$. However, Stolarsky [5] recently has observed that (2) does not hold for all $t>0$. In fact, if

$$t_0=t_0(n) \tag{3}$$

denotes the smallest number such that (2) holds for $t\geq t_0$, Stolarsky has proved that

$$t_0(2)=\frac{1}{2}, \tag{4}$$

$$t_0(n)\leq 1+\frac{1}{n-1}-\frac{1}{\log n}\equiv t_1(n) \quad \text{for } n\geq 3, \tag{5}$$

$$t_0(n)>1-\frac{4\log\log n}{\log 2n} \quad \text{for } n \text{ sufficiently large.} \tag{6}$$

Furthermore, he has conjectured that there are numbers n and t such that (2) is valid and equality holds in (2) for at least one $\mathbf{p}$ other than $(\frac{1}{n},\cdots,\frac{1}{n})$. In the case of Shannon's inequality, the latter probability vector is the only one for which equality holds.

In this paper we shall prove that the best possible exponent t_0 is given by

$$t_0(n)=\sup_{z\in(0,1)} \varphi_n(z), \tag{7}$$

where φ_n is defined for $n\geq 2$, $z\in(0,1)$ by

$$\varphi_n(z)=\frac{\log(n-1)+\log(-\log(1-z))-\log\log(1+(n-1)z)}{\log(1+(n-1)z)-\log(1-z)}. \tag{8}$$

* Received by the editors August 24, 1981.

† Institut für Mathematische Statistik der Westfälischen Wilhelms-Universität Münster, Einsteinstraße 62, 4400 Münster, Germany.

Note that, if we put $\mathbf{p}=((1-z)/n,\cdots,(1-z)/n,(1-z)/n+z)$ for $z\in(0,1)$, then

$$\varphi_n(z)\leq t$$

is equivalent to

$$H_t(\mathbf{p})\leq\log n,$$

as can be easily checked. Therefore, Shannon's inequality yields that φ_n is uniformly bounded by 1, and the supremum of φ_n is obviously a lower bound for $t_0(n)$. An explicit expression for the supremum does not seem to be possible, since the derivative of φ_n is rather unwieldy. Some numerical results and a fairly good approximation for $t_0(n)$ will be given in §4. We shall also see that for $n\geq 3$ and $t=t_0(n)$, $\sup\varphi_n(z)$ is attained for some $z\in(0,1)$, proving that there is a probability vector $\mathbf{p}\neq(\frac{1}{n},\cdots,\frac{1}{n})$ for which equality is obtained in (2). This confirms the conjecture of Stolarsky. The main tool in the proof will be the majorization ordering which we are now going to describe.

2. Majorization. In the sequel, the majorization order will be an essential tool. We therefore briefly recall some notation and facts concerning majorization. For further information we refer to the recent book [2].

For any $\mathbf{x}=(x_1,\cdots,x_n)\in\mathbb{R}^n$, let

$$x_{[1]}\geq\cdots\geq x_{[n]} \tag{9}$$

denote the components of $\mathbf{x}$ in decreasing order. For $\mathbf{x}=(x_1,\cdots,x_n)$ and $\mathbf{y}=(y_1,\cdots,y_n)$ in $\mathbb{R}^n$, $\mathbf{x}$ is said to be *majorized* by $\mathbf{y}$, $\mathbf{x}\prec\mathbf{y}$, if

$$\sum_{i=1}^{k}x_{[i]}\leq\sum_{i=1}^{k}y_{[i]}\qquad(k=1,\cdots,n), \tag{10}$$

with equality holding for $k=n$. For example, if $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, then

$$(\bar{x},\cdots,\bar{x})\prec(x_1,\cdots,x_n). \tag{11}$$

Majorization is equivalent to the following condition:

$$\sum_{i=1}^{n}g(x_i)\leq\sum_{i=1}^{n}g(y_i) \tag{12}$$

for all convex functions $g:\mathbb{R}\to\mathbb{R}$. (See [2, Prop. 4.B.1].) If $\mathbf{y}$ is not a permutation of $\mathbf{x}$ and $\mathbf{x}\prec\mathbf{y}$, then strict inequality holds in (12) if g is strictly convex.

We shall need the following result ([2], Prop. 5.C.1):

LEMMA 1 (Kemperman). *Suppose that $a<b$ and $\mathbf{x}\in[a,b]^n$. Then there exists a unique $c\in[a,b)$ and a unique integer $l\in\{0,\cdots,n\}$ such that*

$$\sum_{i=1}^{n}x_i=(n-l-1)a+c+lb.$$

With c and l so determined,

$$\mathbf{x}\prec\Big(\underbrace{a,\cdots,a}_{n-l-1},c,\underbrace{b,\cdots,b}_{l}\Big). \tag{13}$$

The following consequence of this lemma will be crucial:

PROPOSITION 1. *Let $a\leq x_0\leq b$ and let $g:[a,b]\to\mathbb{R}$ be strictly convex in $[a,x_0]$ and strictly concave in $[x_0,b]$. Then, for every $(x_1,\cdots,x_n)\in[a,b]^n$, there are a nonnegative integer l, $l\leq n$, and numbers $x,y\in[a,b)$, $y\leq x$, such that*

$$(n-l-1)y+x+lb=\sum_{i=1}^{n}x_i$$

and

$$(n-l-1)g(y)+g(x)+lg(b)\leq \sum_{i=1}^{n} g(x_i). \tag{14}$$

Unless

$$(x_1,\cdots,x_n)\doteq\Big(\underbrace{y,\cdots,y}_{n-l-1},x,\underbrace{b,\cdots,b}_{l}\Big)$$

for some y,x, $y\leq x$, *strict inequality can be achieved* ($\doteq$ *denotes equality up to permutation*).

Proof. If $x_0=b$, we can take $l=0$ and $x=y=\frac{1}{n}\sum_{i=1}^{n}x_i$ by virtue of (11) and (12). If $x_0=a$, the proposition follows from Kemperman's lemma. Now assume $a<x_0<b$, and let k be such that $x_{[1]}>x_0,\cdots,x_{[k]}>x_0$ and $x_{[k+1]}\leq x_0$. (If there is no such k, then $x_1,\cdots,x_n\in[a,x_0]$, and the proof follows from (11) and (12) again.)

Since g is concave in $[x_0,b]$ we may apply Kemperman's lemma to $(x_{[1]},\cdots,x_{[k]})\in[x_0,b]^k$. We thus get an integer l and an $x\in[x_0,b)$ such that

$$(k-l-1)x_0+x+lb=\sum_{i=1}^{k}x_{[i]}$$

and

$$(k-l-1)g(x_0)+g(x)+lg(b)\leq \sum_{i=1}^{k} g(x_{[i]}). \tag{15}$$

Now let $y=(1/(n-l-1))(\sum_{i=k+1}^{n}x_{[i]}+(k-l-1)x_0)$. Then

$$\Big(\underbrace{y,\cdots,y}_{n-l-1}\Big)\prec\Big(\underbrace{x_0,\cdots,x_0}_{k-l-1},x_{[k+1]},\cdots,x_{[n]}\Big)$$

by (11), and therefore, using that g is convex in $[a,x_0]$, we obtain

$$(n-l-1)g(y)\leq \sum_{i=k+1}^{n} g(x_{[i]})+(k-l-1)g(x_0). \tag{16}$$

Adding (15) and (16) and cancelling the term $(k-l-1)g(x_0)$ on both sides of the result yields (14). The assumption that g is strictly convex (resp. strictly concave) in $[a,x_0]$ (resp. $[x_0,b]$) guarantees that strict inequality can be obtained in (14) in all nontrivial cases. □

3. Proof of the main result. By putting $x_i=np_i$ we see that inequality (2) is equivalent to

$$\phi_t(\mathbf{x})=\sum_{i=1}^{n} x_i^t\log x_i\geq 0 \tag{17}$$

for all $\mathbf{x}=(x_1,\cdots,x_n)$, $x_i\geq 0$ $(i=1,\cdots,n)$, $\sum_{i=1}^{n}x_i=n$. Let us assume $0<t<1$ since for $t\geq 1$ inequality (2) follows from Shannon's inequality. Let, for $x>0$,

$$g_t(x)=x^t\log x, \tag{18}$$

and put $g_t(0)=0$. Since $g_t''(x)=x^{t-2}(2t-1-t(1-t)\log x)$ for $x>0$, g_t is strictly convex in $[0,x_t]$ and strictly concave in $[x_t,\infty)$, where

$$x_t=\exp\Big(\frac{2t-1}{t(1-t)}\Big). \tag{19}$$

It is worth noting that even without using Kemperman's lemma and Proposition 1, one can give an estimate for t_0 that is almost as good as Stolarsky's but has a much shorter proof:

LEMMA 2. *Let* $n\geq 2$ *and*

$$t_2(n)=\frac{1}{2}-\frac{1}{\log n}+\sqrt{\frac{1}{4}+\frac{1}{(\log n)^2}}\,. \tag{20}$$

For $t\geq t_2(n)$, *inequality* (2) *holds, with equality if and only if* $\mathbf{p}=(\frac{1}{n},\cdots,\frac{1}{n})$.

Proof. $t\geq t_2(n)$ is equivalent to $n\leq x_t$. Since in this case g_t is strictly convex in $[0,n]$, we see from (12) that $\sum_{i=1}^n g_t(x_i)\geq\sum_{i=1}^n g_t(1)=0$, with equality only if $x_i=1$ for all i. □

The exponent $t_2(n)$ is of the same order of magnitude as $t_1(n)$ in (5) but is larger for all n.

We now prove the main result.

THEOREM. *Let* $n\geq 2$. *Then inequality* (2) *holds if and only if*

$$t\geq t_0(n)=\sup_{z\in(0,1)}\varphi_n(z). \tag{21}$$

Equality holds in (2) *if and only if either* $\mathbf{p}=(\frac{1}{n},\cdots,\frac{1}{n})$ *or*

$$t=t_0(n)=\varphi_n(z)$$

and

$$\mathbf{p}\doteq\left(\frac{1-z}{n},\cdots,\frac{1-z}{n},\frac{1-z}{n}+z\right) \tag{22}$$

for some $z\in(0,1)$.

Proof. Let $0<t<1$. By Proposition 1, ϕ_t attains its minimum at some point

$$\mathbf{x}=\Big(\underbrace{y,\cdots,y}_{n-l-1},x,\underbrace{n,\cdots,n}_{l}\Big),$$

where l is a nonnegative integer, $x,y\in[0,n]$, $y\leq x$, and

$$(n-l-1)y+x+ln=n.$$

Of the two possible values $l=1$ and $l=0$, the former implies

$$\phi_t(\mathbf{x})=g_t(n)>0=\phi_t(1,\cdots,1) \tag{23}$$

and hence can be excluded. Therefore,

$$\mathbf{x}=\left(\frac{n-x}{n-1},\cdots,\frac{n-x}{n-1},x\right),$$

where $(n-x)/(n-1)\leq x<n$, or $1\leq x<n$. This yields

$$\phi_t(\mathbf{x})=(n-1)\left(\frac{n-x}{n-1}\right)^t\log\left(\frac{n-x}{n-1}\right)+x^t\log x$$

which, on substituting $z=(x-1)/(n-1)$, becomes

$$\phi_t(\mathbf{x})=(n-1)(1-z)^t\log(1-z)+(1+(n-1)z)^t\log(1+(n-1)z).$$

Thus we have to find the smallest t such that

$$\phi_t(\mathbf{x})\geq 0\quad\text{for all } z\in(0,1).$$

Using the notation of (8), this can be rewritten as

$$t \geq \varphi_n(z) \quad \text{for all } z \in (0,1), \tag{24}$$

which proves the first part of the theorem.

By Proposition 1, equality can hold in (17) only if

$$\mathbf{x} \doteq (1-z, \cdots, 1-z, 1+(n-1)z) \tag{25}$$

for some $z \in [0,1]$. Here $z=1$ can be ruled out by the argument used for (23). The case $z=0$ clearly always yields equality. If $z \neq 0$, that is, $\mathbf{x} \neq (1, \cdots, 1)$, then $\phi_t(\mathbf{x})$ is a strictly increasing function of t; thus in this case equality holds in (17) if and only if $t=\varphi_n(z)$ for some $z \in (0,1)$. This proves (22). □

4. Remarks. From (8) we see that $\varphi_n(z)$ becomes asymptotically equal to $\log\log(1-z)^{-1}/\log(1-z)^{-1}$ as z approaches 1. In particular,

$$\lim_{z \to 1} \varphi_n(z) = 0 \quad \text{and} \quad \lim_{z \to 1} \varphi_n'(z) = -\infty.$$

A somewhat lengthier calculation shows that $\varphi_n(z)$ has the expansion $\varphi_n(z) = \frac{1}{2} + z(n-2)/2 + O(z^2)$ at $z=0$. Thus we have

PROPOSITION 2. *If $n \geq 3$ and $t = t_0(n)$, equality holds in* (2) *for a probability vector different from* $(\frac{1}{n}, \cdots, \frac{1}{n})$.

Proof. Since φ_n is increasing at $z=0$ and decreasing at $z=1$, it attains its supremum somewhere in $(0,1)$. Now use (21) and (22). □

Numerical calculations suggest that φ_n is strictly concave for $n \geq 3$, and hence the equality case of Proposition 2 is unique. We did not succeed in proving this. Except near the endpoints, the graph of φ_n ($n \geq 3$) is almost horizontal. For $n=2$, φ_n is decreasing, so that $t_0(2) = \varphi_2(0) = \frac{1}{2}$.

We conclude by giving in Table 1 a few numerical values, rounded to four decimals. It appears that

$$\varphi_n\left(\frac{2}{5}\right) = \frac{\log(n-1) + \log\log\frac{5}{3} - \log\log\frac{2n+3}{5}}{\log\frac{2n+3}{3}}$$

is a close approximation to $t_0(n)$ for $n < 10^{30}$.

TABLE 1

n	3	4	5	6	10	10^2	10^3	10^4	10^{10}	10^{30}
$t_2(n)$ [see (20)]	.6283	.6563	.6762	.6912	.7280	.8280	.8758	.9031	.9585	.9857
$t_1(n)$ [see (5)]	.5898	.6120	.6287	.6419	.6768	.7930	.8562	.8915	.9566	.9856
$t_0(n) = \max \varphi_n$	.5049	.5119	.5184	.5242	.5416	.6206	.6834	.7297	.8532	.9364
$\varphi_n(0.4)$	.5032	.5115	.5184	.5242	.5414	.6202	.6834	.7295	.8514	.9346

Note added in proof. The author is indebted to J. Aczél for drawing his attention to reference [6]. In §6 of this paper, the type t entropy seems to have been occurred for the first time.

REFERENCES

[1] J. N. KAPUR, *On some properties of generalised entropies*, Indian J. Math., 9 (1967), pp. 427–442.
[2] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[3] P. N. RATHIE, *On some new measures of uncertainty, inaccuracy and information and their characterizations*, Kybernetika, 7 (1971), pp. 394–402.
[4] _____, *Generalized entropies in coding theory*, Metrika, 18 (1972), pp. 216–219.
[5] K. B. STOLARSKY, *A stronger logarithmic inequality suggested by the entropy inequality*, this Journal 11 (1980), pp. 242–247.
[6] J. ACZÉL AND Z. DARÓCZY, *Über verallgemeinerte quasilineare Mittelweret, die mit Gewichtsfunktionen gebildet sind*, Publicationes Math., 10 (1963), pp. 171–190.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0001 $01.25/0

MATRIX STIELTJES SERIES AND NETWORK MODELS*

S. BASU† AND N. K. BOSE‡

Abstract. Simple proofs are given of the fact that Padé approximants of certain orders, to a symmetric matrix power series, are realizable as RC multiports, possibly containing ideal transformers. One proof technique uses the matrix continued fraction expansion while the other uses the theory of matrix Cauchy index. An important offshoot of the main result is the presentation of properties of a set of matrix polynomials, which are orthogonal over a real semi-infinite interval.

1. Introduction. The fact that scalar Padé approximants of certain orders to a Stieltjes power series characterize RC driving point functions is well known [1]. Very recently, the matrix counterpart of this result has been given [2]. The primary objective of this paper is to provide alternate proofs of this result, via the simple artifices of matrix continued fraction [3] and the more recently developed tool of matrix Cauchy index [4]. Furthermore, it is shown that the "denominator" polynomial matrices of the Padé approximants to a matrix Stieltjes series form a sequence of polynomial matrices orthogonal over a real semi-infinite interval. Matrix counterparts of many results related to the classical orthogonal polynomials are derived and, in the context of present discussion, these results also fall out as consequences of network theoretic interpretations given to the matrix Padé approximants. It may be noted that matrix extensions of the theory of polynomial matrices orthogonal on the unit circle has been carried out recently [5], [10]

In §2 of this paper, the matrix Stieltjes series is introduced and the equivalence between the Padé approximants to this series and the matrix continued fraction of Shieh et al. [3] with positive definite coefficients is established, thereby providing a proof, via synthesis, of multiport impedance realizability of the approximants. In §3 matrix extensions of some results related to the classical theory of orthogonal polynomials are presented. These results, coupled with those on matrix Cauchy index, then make feasible the presentation in §4 of a more direct discussion of multiport RC realizability of Padé approximants to a matrix Stieltjes series. Finally, in §5, it is pointed out that some of the results proved in §4 follow directly from the network interpretations given to a matrix Stieltjes series.

2. Matrix Stieltjes series and RC continued fraction synthesis of Padé approximants. Consider a formal matrix power series,

$$T(s)=\sum_{k=0}^{\infty} T_k s^k, \tag{2.1}$$

where T_k is a symmetric $(p\times p)$ matrix. Square block Hankel matrices, $H_n(T)$ and $H_n'(T)$, are associated with the series $T(s)$ in (2.2) below.

* Received by the editors February 4, 1981, and in revised form August 25, 1981. This research was supported by the U.S. Air Force Office for Scientific Research under grant 78-3542 and the National Science Foundation under grant 78-23141.

† Department of Electrical Engineering, Stevens Institute of Technology, Hoboken, New Jersey 07030.

‡ Departments of Electrical Engineering and Mathematics, 348 Benedum Hall, University of Pittsburgh, Pittsburgh, PA 15261.

(2.2)

$$H_n(T)=\begin{bmatrix} T_0 & T_1 & T_2 & \cdots & T_n \\ T_1 & T_2 & & & \\ T_2 & T_3 & & & \vdots \\ \vdots & & & & \\ T_n & & \cdots & & T_{2n} \end{bmatrix}, \quad H_n'(T)=\begin{bmatrix} T_1 & T_2 & T_3 & \cdots & T_n \\ T_2 & T_3 & & & \\ T_3 & T_4 & & & \vdots \\ \vdots & & & & \\ T_n & & \cdots & & T_{2n-1} \end{bmatrix}$$

DEFINITION 1. The series $T(s)$ in (2.1) will be called a matrix Stieltjes series if $H_n(T)$ is positive definite and $H_n'(T)$ is negative definite for $n=0,1,2,\cdots$. The following preliminary results will be of use in later discussions.

LEMMA 2.1. *If $T(s)$ is any symmetric formal matrix power series as in* (2.1), *then its formal inverse denoted by $[T(s)]^{-1}$ exists, is unique and it is also symmetric.*

Proof. First it will be shown that $[T(s)]^{-1}=\Sigma_{k=0}^{\infty} W_k s^k$ exists such that

$$\left(\sum_{k=0}^{\infty} T_k \cdot s^k\right)\left(\sum_{k=0}^{\infty} W_k \cdot s^k\right)=1. \tag{2.3}$$

Equation (2.3) would imply that (2.4) has to hold for any n.

$$F(T)=\begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix}=\begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.4}$$

where $F(T)$ is the lower triangular block Toeplitz matrix in (2.5) associated with the power series $T(s)$ in (2.1).

$$F(T)=\begin{bmatrix} T_0 & & & \\ T_1 & & 0 & \\ \vdots & \ddots & & \\ T_n & \cdots & T_1 & T_0 \end{bmatrix} \tag{2.5}$$

Positive definiteness of T_0, then, implies the existence of matrices W_k, $k=0,1,\cdots$. Uniqueness follows from the invertibility of T_0. The fact that W_k's are symmetric is proved next. Since the elements of $T(s)$ belong to an integral domain,

$$\left[(T(s))^{-1}\right]^t=\left[(T(s))^t\right]^{-1}=[T(s)]^{-1}. \tag{2.6}$$

The last equality follows from the symmetry of T_k's and the proof of the lemma is now complete.

LEMMA 2.2. *If $T(s)$ is a matrix Stieltjes series, then so is $G(s)$, given by*

$$T(s)=[A_1+sG(s)]^{-1}, \tag{2.7}$$

where $A_1=[T(0)]^{-1}$.

Proof. The proof will be in two different parts. In the first part a proof for the special case $T_0=I=A_1$, where I is an identity matrix of appropriate order, will be provided, and in the second part it will be shown that no loss of generality occurs under this assumption.

Part I. Equation (2.8) follows from (2.7).

$$T(s)[I+sG(s)]=[I+sG(s)]T(s)=I. \tag{2.8}$$

Let $G(s)=\sum_{k=0}^{\infty}G_k s^k$, where symmetry of each G_k follows from Lemma 2.1. Associate square block Hankel matrices $H_n(G)$ and $H_n'(G)$ of orders $(n+1)p$ and np, as in (2.2). On equating coefficients of like powers of s in (2.8) and defining

$$S_n=\begin{bmatrix} I & & & \\ G_0 & I & & 0 \\ \vdots & \ddots & \ddots & \\ G_{n-1} & \cdots & G_0 & I \end{bmatrix}, \tag{2.9}$$

(2.10) and (2.11) can be easily verified.

$$S_{n+1}H_n(T)S_{n+1}^t=\left[\begin{array}{c|c} I & 0 \\ \hline 0 & -H_n'(G) \end{array}\right], \tag{2.10}$$

$$S_{n+1}H_{n+1}'(T)S_{n+1}^t=-H_n(G). \tag{2.11}$$

In (2.10), positive definiteness of $H_n(T)$ implies negative definiteness of $H_n'(G)$, and in (2.11) negative definiteness of $H_{n+1}'(T)$ implies positive definiteness of $H_n(G)$.

Part 2. If $T_0\neq I$ then consider the series $T'(s)=T_0^{-1/2}\cdot T(s)\cdot T_0^{-1/2}$, where $T_0^{1/2}$ is the Hermitian square root of T_0 and $T_0^{-1/2}$ is its inverse. Next, it is noted that if $H_n(T')$ and $H_n'(T')$ are the block Hankel matrices as in (2.2) associated with the series $T'(s)$ then (2.12) and (2.13) holds, implying that, when $T(s)$ is a matrix Stieltjes series, $H_n(T')$ and $H_n'(T')$ are positive definite and negative definite respectively.

$$H_n(T')=\operatorname{diag}\left(T_0^{-1/2}T_0^{-1/2}\cdots T_0^{-1/2}\right)H_n(T)\operatorname{diag}\left(T_0^{-1/2}T_0^{-1/2}\cdots T_0^{-1/2}\right) \tag{2.12}$$

and

$$H_n'(T')=\operatorname{diag}\left(T_0^{-1/2}T_0^{-1/2}\cdots T_0^{-1/2}\right)H_n'(T)\operatorname{diag}\left(T_0^{-1/2}T_0^{-1/2}\cdots T_0^{-1/2}\right). \tag{2.13}$$

Thus the series $T'(s)$ is also a matrix Stieltjes series. Using Part 1 above gives $T'(s)=[A_1'+s.G'(s)]^{-1}$, where $A_1'=I$ is positive definite and $G'(s)$ is a matrix Stieltjes series. Therefore, $T(s)=T_0^{1/2}.T'(s).T_0^{1/2}=[A_1+s.G(s)]^{-1}$, with $A_1=T_0^{-1/2}.A_1'.T_0^{-1/2}$ and $G(s)=T_0^{-1/2}.G'(s).T_0^{-1/2}$. Positive definiteness of A_1 is obvious from positive definiteness of A_1'. Finally, invoking the type of argument used in connection with (2.12) and (2.13), it readily follows that $G(s)$ is a matrix Stieltjes series. The proof of the lemma is therefore, complete.

Repeated use of Lemma 2.2 yields the following expansion for a matrix Stieltjes series $T(s)$:

$$T(s)=\left[A_1+\left[\frac{1}{s}A_2+\cdots\left[\frac{1}{s^{l(k)}}A_k+s^{-l(k)+1}T_k(s)\right]^{-1}\cdots\right]^{-1}\right]^{-1}, \tag{2.14}$$

where $l(k)=0$ for k odd and $l(k)=1$ for k even. A_j, for $j=1,2,\cdots,k$, are symmetric positive definite matrices and $T_k(s)$ is a matrix Stieltjes series. It is interesting to note that the matrix continued fraction in (2.14) is of the same type as has been dealt with in [3].

The proof of the scalar version of the following lemma is given in [1, p. 55], [6, p. 380]. A proof for the matrix case can be constructed exactly along parallel lines and is omitted for the sake of brevity.

LEMMA 2.3. *For each* $k=1,2,3,\cdots$, *the kth convergent to the matrix continued fraction in* (2.14) *is equal to the matrix Padé approximant of order* $[(m-1)/m]$ *or* $[m/m]$ *to* $T(s)$, *according as* $k=2m$ *or* $k=2m+1$.

In the topic of passive reciprocal network synthesis, elements like *resistors*, *inductors*, *capacitors* and transformers are the fundamental building blocks. Rational matrices which characterize the input-output behaviour of such networks have distinguishing properties and a detailed documentation of these and related materials is available in [7]. Networks modeled from resistors, capacitors and ideal transformers form an important subclass of the class of passive, reciprocal networks. A rational matrix with special properties can be synthesized in various ways to yield different networks, each of whose input-output behaviour is characterizable by the specified rational matrix. One synthesis technique is based on the theory of continued fraction expansion. In particular, for the kth convergent associated with (2.14), the positive definiteness of the coefficients matrices A_i, $i=1,2,\cdots,k$ guarantees the feasibility of synthesis of the convergents using resistors, capacitors and ideal transformers (RC-ideal transformer) [7, pp. 216–218] as circuit elements. It then becomes clear from the results developed so far that a RC-ideal transformer realization of Padé approximants of orders $[(m-1)/m]$ and $[m/m]$ to $T(s)$ in (2.14) is possible. This conclusion linking the continued-fraction expansion associated with a matrix Stieltjes series and the important topic of RC network synthesis (in microminiaturized circuits, the elements fabricated in integrated form are resistors, capacitors and active elements like transistors) is summarized in Theorem 2.1 below.

THEOREM 2.1. *The* $[(m-1)/m]$ *and* $[m/m]$ *order Padé approximants to a matrix Stieltjes series can be synthesized as multiports using resistors and capacitors (possibly including ideal transformers).*

3. Matrix analogues of orthogonal polynomials on real intervals. It is known [1] that the sequence of denominator polynomials of scalar Padé approximants to a Stieltjes series forms an orthogonal polynomial sequence, orthogonal on the real interval $(-\infty,0]$. In the present section an investigation into whether or not similar results also hold in the matrix case is carried out. Matrix counterparts of many classical results related to orthogonal polynomials on a real interval is obtained.

To begin with, some notations are introduced first. Consider the set of linear simultaneous equations in (3.1):

$$\begin{bmatrix} T_0 & T_1 & T_2 & \cdots & T_n \\ T_1 & T_2 & & & \\ T_2 & T_3 & & & \vdots \\ \vdots & & & & \\ T_n & & & \cdots & T_{2n} \\ T_{n+1} & & & \cdots & T_{2n+1} \end{bmatrix} \begin{bmatrix} p_n^{(n)} \\ p_{n-1}^{(n)} \\ \vdots \\ \\ p_1^{(n)} \\ I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ D_n \\ E_n \end{bmatrix}, \tag{3.1}$$

where $p_k^{(n)}$, $k=1,2,3,\cdots$, and D_n are $(p\times p)$ matrices, I is an identity matrix and 0 is a null matrix, each of order $(p\times p)$. Then the polynomial matrix $P_n(s)$ in (3.2) is the

$$P_n(s)=\sum_{k=0}^{n} p_k^{(n)}s^k, \qquad p_0^{(n)}=I \tag{3.2}$$

"denominator" polynomial matrix associated with the Padé aproximant of order $[(n-1)/n]$ to the series $T(s)$. Note that in (3.1) and (3.2) the superscript (n) is taken to

mean that Padé approximants of order $[(n-1)/n]$ are being considered. Consider, also, the "inverse polynomial matrix," $P_n'(s)$ in (3.3).

$$P_n'(s)=\sum_{k=0}^{n} p_{n-k}^{(n)} s^k. \tag{3.3}$$

It will be seen in what follows that under the assumption that $H_n(T)$ is positive definite for all n, the polynomial matrix $P_n'(s)$ enjoys properties which can be considered as matrix counterparts of the properties of classical scalar orthogonal polynomials. Before taking up a detailed discussion of the properties, a few results which are almost obvious are presented in Lemma 3.1. To avoid clutter in notation the matrices K_n and M_n are introduced.

$$K_n \triangleq E_n D_n^{-1} - E_{n-1} D_{n-1}^{-1}, \tag{3.4}$$

$$M_n = \begin{bmatrix} p_n^{(n)t} & p_{n-1}^{(n)t} & \cdots & p_1^{(n)t} & I \end{bmatrix}^t. \tag{3.5}$$

LEMMA 3.1. *If $H_n(T)$ is positive definite for all n, then*: (i) *D_n is symmetric positive definite*, (ii) *$K_n D_n$ is symmetric.*

Proof. (i) From (3.1) and (3.5), $M_n^t H_n(T) M_n = D_n$. The result then follows from the fact that $H_n(T)$ is positive definite and M_n, by definition, is of full column rank. (ii) Again, (3.1) for n and $n-1$ together with (3.5), lead to (3.6).

$$\left[\begin{array}{c|c} M_{n-1}^t & 0 \\ \hline & M_n^t \end{array}\right] H_n(T) \left[\begin{array}{c|c} M_{n-1} & M_n \\ \hline 0 & \end{array}\right] = \begin{bmatrix} D_{n-1} & 0 \\ p_1^{(n)t} D_{n-1} + E_{n-1} & D_n \end{bmatrix}. \tag{3.6}$$

Since the left-hand side of (3.6) is symmetric,

$$p_1^{(n)t} D_{n-1} + E_{n-1} = 0. \tag{3.7}$$

Furthermore, it follows on straightforward premultiplication of (3.1) (excluding the first block row), by the transpose of (3.5) and subsequent use of (3.7) and (3.4),

$$M_n^t H_{n+1}'(T) M_n = p_1^{(n)t} D_n + E_n = -E_{n-1} D_{n-1}^{-1} D_n + E_n = K_n D_n. \tag{3.8}$$

The last two equalities follow via the use of (3.7) and (3.4). The proof of the lemma is thus complete.

Consider now two "inverse polynomial matrices" $P_n'(s)$ and $P_m'(s)$, with $n \geq m$. Define the inner product,

$$\langle P_n'(s), P_m'(s) \rangle = \begin{bmatrix} M_m^t & 0 & \cdots & 0 \end{bmatrix} H_n(T) M_n, \tag{3.9a}$$

where M_n, $H_n(T)$ are defined in (3.5) and (2.2), respectively, while $[M_m^t \; 0 \; \cdots \; 0]$ is formed by augmenting M_m^t (where M_m is defined analogously to M_n in (3.5)) by $(n-m)$ null matrices, each of size $(p \times p)$. It is readily verifiable via use of (3.1) that

$$\langle P_n'(s), P_m'(s) \rangle = \begin{cases} 0, & n \neq m, \\ D_n, & n = m. \end{cases} \tag{3.9b}$$

Note that D_n is symmetric positive definite, as proved in part (i) of Lemma 3.1. In light of the preceding discussion, it is justifiable to say that the set $\{P_k'(s)\}$ of "inverse polynomial matrices" associated with the Padé approximants to a matrix Stieltjes series, forms a sequence of orthogonal polynomial matrices.

To continue the discussion of properties of $P_n'(s)$, under the assumption that $H_n(T)$ is positive definite for all n, note that the grouping of equations determining the

"denominator polynomial matrices" of Padé approximants of orders $[(n-2)/(n-1)]$, $[(n-1)/n]$ leads to (3.10).

$$H_{n+1}(T)\begin{bmatrix} M_{n-1} & \cdot & M_n & \cdot & 0 \\ 0 & \cdot & & \cdot & M_n \\ 0 & \cdot & 0 & \cdot & \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ D_{n-1} & 0 & D_n \\ E_{n-1} & D_n & E_n \\ F_{n-1} & E_n & F_n \end{bmatrix}. \tag{3.10}$$

Right multiplying (3.10) by the matrix N_n in (3.11) yields (3.12). (Note that nonsingularity of D_n and D_{n-1} follows from positive definiteness of $H_n(T)$, $n=0,1,2,\cdots$.)

$$N_n=\left[-D_{n-1}^{-1}D_n : -D_n^{-1}K_nD_n : I\right], \tag{3.11}$$

$$H_{n+1}(T)\begin{bmatrix} M_{n-1} & \cdot & M_n & \cdot & 0 \\ 0 & \cdot & & \cdot & \\ & \cdot & & \cdot & \\ 0 & \cdot & 0 & \cdot & M_n \end{bmatrix} N_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ D_{n+1} \end{bmatrix}, \tag{3.12}$$

where D_{n+1} is naturally defined.

Therefore, in view of (3.1) and (3.12), the recurring relation (3.13) follows.

$$M_{n+1}=\begin{bmatrix} M_{n-1} & \cdot & M_n & \cdot & 0 \\ 0 & \cdot & & \cdot & \\ 0 & \cdot & 0 & \cdot & M_n \end{bmatrix} N_n. \tag{3.13}$$

Finally, noting the relation between M_n and the coefficients of s^k, $k=0,1,\cdots,n$, in $P_n'(s)$, (3.13) yields in (3.14) and (3.15) the matrix version of the familiar three term recurrence formula relating sequence of polynomials orthogonal on a real interval.

$$P'_{k+1}(s)=P'_k(s)(sI-C_k)-P'_{k-1}(s)\lambda_k, \qquad k=1,2,\cdots, \tag{3.14a}$$

$$P'_{-1}(s)=0, \qquad P'_0(s)=I, \tag{3.14b}$$

with

$$C_k=D_k^{-1}K_kD_k, \qquad C_0=D_0^{-1}T_1, \tag{3.15a}$$

$$\lambda_k=D_{k-1}^{-1}D_k. \tag{3.15b}$$

Now, the first major result in this section will be stated and proved.

THEOREM 3.1. *If $H_n(T)$ is positive definite for any integer $n>0$, then*: (i) *the zeros of $|P_n'(s)|$ (i.e., the determinant of $P_n'(s)$) are real*; (ii) *if α_j is a zero of $|P_n'(s)|$ of multiplicity m there exists a set of exactly m linearly independent sets of $(1\times p)$ vectors $\{v_j^1\ v_j^2\cdots v_j^m\}$ such that $v_j^iP_n(\alpha_j)=0$, $i=1,2,\cdots,m$.*

Proof. (i) For any zero $s=\alpha_j$ of $|P_n'(s)|$, there exists at least one $(1\times p)$ vector v such that $vP_n'(\alpha_j)=0$. Next, considering (3.14) with $s=\alpha_j$ for $k=0,1,\cdots,n$ and noting that $vP_n'(\alpha_j)=0$,

$$\left[vP_0'(\alpha_j) \quad vP_1'(\alpha_j) \quad \cdots \quad vP_{n-1}'(\alpha_j)\right]A_n=\left[vP_0'(\alpha_j) \quad vP_j'(\alpha_j) \quad \cdots \quad vP_{n-1}'(\alpha_j)\right]\alpha_j, \tag{3.16}$$

where A_n is the block tridiagonal matrix in familiar form:

$$A_n=\begin{bmatrix} C_0 & \lambda_1 & & & \\ I & C_1 & \lambda_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & & & \lambda_{n-1} \\ 0 & & & I & C_{n-1} \end{bmatrix} \tag{3.17}$$

$$=\operatorname{diag}\begin{bmatrix} D_0^{-1} & D_1^{-1} & \cdots & D_{n-1}^{-1} \end{bmatrix}\begin{bmatrix} T_1 & D_1 & & & 0 \\ D_1 & K_1D_1 & \ddots & & \\ & \ddots & \ddots & & D_{n-1} \\ 0 & & & D_{n-1} & K_{n-1}D_{n-1} \end{bmatrix} \tag{3.18}$$

Equation (3.18) follows from (3.17) via the use of (3.15). Note from (3.16) that α_j is an eigenvalue of A_n. Also in (3.18), in view of Lemma 3.1, the diagonal matrix on the left is positive definite and the tridiagonal matrix on the right is symmetric. Next, invoking a standard result on eigenvalues of quadratic forms [8, p. 310] it is found that α_j is always real.

(ii) To prove the second part of the lemma, it is further noted that if A_n in (3.16) has an eigenvalue α_j of multiplicity m, then there must exist a set of vectors $\{v_j^1\ v_j^2\cdots v_j^m\}$ such that the $(1\times np)$ vectors $[v_j^kP_0'(\alpha_j)\ v_j^kP_1'(\alpha_j)\cdots v_j^kP_{n-1}'(\alpha_j)]$ for $k=1,2,3,\cdots,m$ are all linearly independent. However, linear independence of $[v_j^kP_0'(\alpha_j)\ v_j^kP_1'(\alpha_j)\ \cdots\ v_j^kP_{n-1}'(\alpha_j)]$ implies the linear independence of the set $\{v_j^1\ v_j^2\cdots v_j^m\}$. The proof of the lemma is thus complete.

The following two corollaries are immediate consequences of the above theorem.

COROLLARY 3.1. *Any zero of $|P_n'(s)|$ cannot be of multiplicity larger than p.*

Proof. The above result is obvious from the fact that if $s=\alpha$ is any zero of $|P_n'(s)|=0$ then $vP_n'(\alpha)=0$ cannot have more than p linearly independent solutions for v.

COROLLARY 3.2. *The invariant factors in the Smith canonical form for $P_n'(s)$ cannot have zeros of multiple order.*

Proof. Let $v_1(s),v_2(s),\cdots,v_p(s)$ denote the invariant factors. Obviously $v_j(\alpha)=0$ for some j implies $|P_n'(\alpha)|=0$. Also if α is of multiplicity m, invoking part (ii) of the above theorem yields that $P_n'(\alpha)$ is of rank $(p-m)$, which in turn implies that $v_j(\alpha)=0$ for exactly m different values of j. The required result then follows by noting the fact that the multiplicity of factor $(s-\alpha)$, in $|P_n'(s)|=\Pi_{j=1}^{p}v_j(s)$, is exactly equal to m.

Interestingly, it is noted that Corollaries 3.1 and 3.2 can be considered as the matrix interpretation of the scalar result that zeros of orthogonal polynomials are not only real but also simple.

Before proceeding further a polynomial matrix $K_n(s,u)$ in two variables s and u is introduced.

$$K_n(s,u)=\sum_{k=0}^{n}P_k'(s)D_k^{-1}P_k''(u). \tag{3.19}$$

In keeping with the scalar theory $K_n(s,u)$ will be called the kernel polynomial matrix. The following result is the matrix version of Christoffel-Darboux identity [11].

THEOREM 3.2. *The following is true for all values of n.*

$$(s-u)K_n(s,u)=P_{n+1}'(s)D_n^{-1}P_n''(u)-P_n'(s)D_n^{-1}P_{n+1}''(u). \tag{3.20}$$

The proof for the above theorem runs exactly parallel to the scalar case and is outlined in the following. Postmultiplying (3.14a) by $D_k^{-1}P_k''(u)$, $k>0$, and rearranging

terms, after using (3.15) one obtains

$$sP'_k(s)D_k^{-1}P''^t_k(u)=P'_{k+1}(s)D_k^{-1}P''^t_k(u)+P'_k(s)D_k^{-1}K_kP''^t_k(u) \\ +P'_{k-1}(s)D_{k-1}^{-1}P''^t_k(u). \tag{3.21}$$

Again taking transpose of (3.14a), with s replaced by the variable u, premultiplying by $P'_k(s)D_k^{-1}$ and rearranging, after using (3.15), one obtains

$$uP'_k(s)D_k^{-1}P''^t_k(u)=P'_k(s)D_k^{-1}P''^t_{k+1}(u)+P'_k(s)K_k^tD_k^{-1}P''^t_k(u) \\ +P'_k(s)D_{k-1}^{-1}P''^t_{k-1}(u). \tag{3.22}$$

Next, subtracting (3.22) from (3.21), summing the resulting equation through $k=1,2,\cdots,n$ and adding to the resulting sum the corresponding equation for $k=0$ (use (3.14) with k set to zero), the desired equality is obtained (use has to be made of the fact that K_kD_k is symmetric). The final property require will be a matrix version of the Gauss quadrature formula. In the scalar case, at least two different forms of the formula exist in the literature. The matrix version of the form given by Gragg [9] will be proved here.

Let α_j be a zero of $|P'_n(s)|$ of multiplicity m_j. Then, from Theorem 3.1, there exists a linearly independent set of $(1\times p)$ vectors $\{v_j^k\}$, $k=1,2,\cdots,m_j$ such that $v_j^kP'_n(\alpha_j)=0$ for $k=1,2,\cdots,m_j$. Next, consider the kernel polynomial matrix evaluated at α_j

$$K_n(\alpha_j,\alpha_j)=D_0^{-1}+\sum_{k=1}^{n}P'_k(\alpha_j)D_k^{-1}P''^t_k(\alpha_j).$$

Since D_i is positive definite, $K_n(\alpha_j,\alpha_j)$ is a positive definite symmetric matrix. Therefore $K_n(\alpha_j,\alpha_j)$ can be taken to induce an inner product in the vector space spanned by the linearly independent set $\{v_j^k\}$, $k=1,2,\cdots,m_j$. If $\{u_j^1,u_j^2,\cdots,u_j^{m_j}\}$ is a set of orthogonal vectors in this space obtained from $\{v_j^1\ v_j^2\cdots v_j^{m_j}\}$, by means of Gram–Schmidt orthogonalization procedure, then

$$u_j^\kappa\cdot K_n(\alpha_j,\alpha_j)\big(u_j^\mu\big)^t=\begin{cases}0 & \text{for } \kappa\neq\mu,\\ 1/G_{n_j}^\kappa & \text{for } \kappa=\mu,\end{cases} \tag{3.23}$$

where, $G_{n_j}^\kappa$ is a real positive number.

Let it be assumed that the above construction has been carried out for each zero $\alpha_j, j=1,2,\cdots,l$ of $|P'_n(s)|$, where l is the number of distinct zeros of $|P'_n(s)|$.

Two notations are introduced next, to denote the matrices U_n and V_n of sizes $(m\times mp)$ and $(mp\times np)$, respectively, where $(m_1+m_2+\cdots+m_l)\triangleq m$.

$$U_n\triangleq\operatorname{diag}\big[u_1^1\ \cdots\ u_1^{m_1}\ u_2^1\ \cdots\ u_2^{m_2}\ \cdots\ u_l^1\ \cdots\ u_l^{m_l}\big], \tag{3.24}$$

$$V_n\triangleq\left[\begin{array}{ccccc}
I & I\alpha_1 & I\alpha_1^2 & \cdots & I\alpha_1^{n-1}\\
\vdots & & & & \vdots\\
I & I\alpha_1 & I\alpha_1^2 & \cdots & I\alpha_1^{n-1}\\
\hline
\vdots & & & & \vdots\\
\hline
I & I\alpha_l & I\alpha_l^1 & \cdots & I\alpha_l^{n-1}\\
\vdots & & & & \vdots\\
I & I\alpha_l & I_l & \cdots & I\alpha_l^{n-1}
\end{array}\right]\begin{array}{l}\\ \\ m_1\text{ block rows}\\ \\ \\ \\ m_l\text{ block rows}\end{array}. \tag{3.25}$$

The following results will be the key to the matrix Gauss quadrature formula.

LEMMA 3.2. *The following hold true*:

(i)

$$u_j^{\kappa} K_{n-1}(\alpha_j, \alpha_j)\big(u_j^{\mu}\big)^t = \begin{cases} 0 & \text{for } \kappa \neq \mu, \\ 1/G_{n_j}^{\kappa} & \text{for } \kappa = \mu \end{cases}$$

(ii)

$$u_i^{\kappa} K_{n-1}(\alpha_i, \alpha_j)\big(u_j^{\mu}\big)^t = 0 \text{ for } i \neq j \text{ and any value of } \kappa \text{ and } \mu.$$

(iii) *The* $(m \times np)$ *matrix* $U_n V_n$ *is of full rank.*

Proof: (i) By construction of $\{u_j^{\kappa}\}$, $\kappa = 1, 2, \cdots, m_j$.

(ii) From Theorem 3.2 it follows that

$$(\alpha_i - \alpha_j) K_n(\alpha_i, \alpha_j) = P'_{n+1}(\alpha_i) D_n^{-1} P''_n(\alpha_j) - P'_n(\alpha_i) D_n^{-1} P''_{n+1}(\alpha_j). \tag{3.26}$$

Then the required result easily follows by noting that $u_i^{\kappa} P'_n(\alpha_i) = 0$,

$$u_j^{\mu} \cdot P'_n(\alpha_j) = 0 \quad \text{and} \quad u_i^{\kappa} K_n(\alpha_i, \alpha_j)\big(u_j^{\mu}\big)^t = u_i^{\kappa} K_{n-1}(\alpha_i, \alpha_j)\big(u_j^{\mu}\big)^t.$$

(iii) The following notation for the $(np \times np)$ upper triangular block Toeplitz matrix in (3.27) is needed

$$L_n \triangleq \begin{bmatrix} I & p_1^{(1)} & p_2^{(2)} & \cdots & p_{n-1}^{(n-1)} \\ & I & p_1^{(2)} & \cdots & p_{n-2}^{(n-1)} \\ & & \vdots & \cdots & \vdots \\ & & I & \cdots & p_2^{(n-1)} \\ 0 & & & \cdots & p_1^{(n-1)} \\ & & & & I \end{bmatrix}. \tag{3.27}$$

Note that on carrying out straight forward block matrix multiplication

$$(V_n L_n) = \left[\begin{array}{cccc} P'_0(\alpha_1) & P'_1(\alpha_1) & \cdots & P'_{n-1}(\alpha_1) \\ \vdots & \vdots & & \vdots \\ P'_0(\alpha_1) & P'_1(\alpha_1) & \cdots & P'_{n-1}(\alpha_1) \\ \hline \vdots & & & \vdots \\ \hline P'_0(\alpha_l) & P'_1(\alpha_l) & \cdots & P'_{n-1}(\alpha_l) \\ \vdots & \vdots & & \vdots \\ P'_0(\alpha_l) & P'_1(\alpha_l) & \cdots & P'_{n-1}(\alpha_l) \end{array}\right] \begin{array}{l} m_1 \text{ block rows} \\ \\ \\ \hline \\ \hline \\ m_l \text{ block rows} \end{array}. \tag{3.28}$$

Next, using the notation of (3.27), it follows from (3.1) in a compact form:

$$L_n^t H_{n-1}(T) L_n = \operatorname{diag}(D_0 \; D_1 \; \cdots \; D_{n-1}). \tag{3.29}$$

Taking the inverse of (3.29), transferring the L_n on the right-hand side and them premultiplying and postmultiplying the resulting equation by $U_n V_n$ and $(U_n V_n)^t$, respectively, (3.30) is obtained.

$$(3.30)\quad (U_nV_n)H_{n-1}^{-1}(T)(U_nV_n)^t = U_n\Big[(V_nL_n)\big\{\mathrm{diag}\big(D_0^{-1}\,D_1^{-1}\,\cdots\,D_{n-1}^{-1}\big)\big\}(V_nL_n)^t\Big]U_n^t.$$

Carrying out the matrix multiplication and using (3.19) the quantity inside square bracket in the right-hand side becomes:

(3.31)

$$(V_nL_n)\big\{\mathrm{diag}\big(D_0^{-1}\,D_1^{-1}\,\cdots\,D_{n-1}^{-1}\big)\big\}(V_nL_n)^t$$

$$= \left[\begin{array}{ccc|c|ccc}
K_{n-1}(\alpha_1,\alpha_1) & \cdots & K_{n-1}(\alpha_1,\alpha_1) & \cdots & K_{n-1}(\alpha_1,\alpha_l) & \cdots & K_{n-1}(\alpha_1,\alpha_l) \\
\vdots & & \vdots & \cdots & \vdots & & \vdots \\
K_{n-1}(\alpha_1,\alpha_1) & \cdots & K_{n-1}(\alpha_1,\alpha_1) & \cdots & K_{n-1}(\alpha_1,\alpha_l) & \cdots & K_{n-1}(\alpha_1,\alpha_l) \\
\hline
 & \cdots & & \cdots & & \cdots & \\
 & \cdots & & & & \cdots & \\
\hline
K_{n-1}(\alpha_l,\alpha_1) & \cdots & K_{n-1}(\alpha_l,\alpha_1) & \cdots & K_{n-1}(\alpha_l,\alpha_l) & \cdots & K_{n-1}(\alpha_l,\alpha_l) \\
\vdots & & \vdots & & \vdots & & \vdots \\
K_{n-1}(\alpha_l,\alpha_1) & \cdots & K_{n-1}(\alpha_l,\alpha_l) & \cdots & K_{n-1}(\alpha_l,\alpha_l) & \cdots & K_{n-1}(\alpha_l,\alpha_l)
\end{array}\right]$$

Next, pre- and postmultiplying the matrix in (3.31) by U_n and U_n^t respectively and taking Lemma 3.2 into account it follows from (3.30) that

$$(3.32)\quad (U_nV_n)H_{n-1}^{-1}(T)(U_nV_n)^t = \Big[\mathrm{diag}\big(G_{n_1}^1\,\cdots\,G_{n_1}^{m_1}\,\vdots\,\cdots\,\vdots\,G_{n_l}^1\,\cdots\,G_{n_l}^{m_l}\big)\Big]^{-1}.$$

Since in (3.32) $H_{n-1}(T)$ and the right-hand side are nonsingular matrices, it follows that (U_nV_n) is of full rank. The proof of the lemma is, therefore, completed.

Finally to arrive at the matrix version of Gauss quadrature formula, (3.32) is written as

$$(3.33)\qquad H_{n-1}(T) = (U_nV_n)^t\mathrm{diag}\big(G_{n_1}^1\,\cdots\,G_{n_1}^{m_1}\,\cdots\,G_{n_l}^1\,\cdots\,G_{n_l}^{m_l}\big)(U_nV_n).$$

Comparing each $(p\times p)$ block element on both sides of (3.33) it follows that

$$(3.34)\qquad T_k = \sum_{j=1}^{l}\sum_{\kappa=1}^{m_j} G_{n_j}^{\kappa}\big(u_j^{\kappa t}\cdot u_j^{\kappa}\big)\cdot\alpha_j^k$$

for $k=0,1,2,\cdots,(2n-2)$.

Also, considering (3.1) again,

$$(3.35a)\qquad T_{2n-1} = -\sum_{r=1}^{n} T_{2n-r-1}\,p_r^{(n)}$$

$$(3.35b)\qquad = -\sum_{r=1}^{n}\sum_{j=1}^{l}\sum_{\kappa=1}^{m_j} G_{n_j}^{\kappa}\big(u_j^{\kappa t}\cdot u_j^{\kappa}\big)\alpha_j^{2n-r-1}p_r^{(n)}$$

(on substituting for T_{2n-r-1} from (3.34)).

Again, since $u_j^{\kappa}P_n^{(r)}(\alpha_j)=0$ making use of (3.3), it follows that $u_j^{\kappa}\Sigma_{r=1}^{n}p_r^{(n)}\alpha_j^{n-r} = -u_j^{\kappa}\cdot\alpha_j^n I$. Interchanging summation in (3.35b) and making use of the above equation,

$$(3.36)\qquad T_{2n-1} = \sum_{j=1}^{l}\sum_{\kappa=1}^{m_j} G_{n_j}^{\kappa}\big(u_j^{\kappa t}\cdot u_j^{\kappa}\big)\alpha_j^{2n-1}.$$

To summarize our result, we write matricial Gauss quadrature formula in the following theorem.

THEOREM 3.3. *If* α_j, $j=1,2,\cdots,l$, *is a zero of* $|P_n'(s)|$, *of multiplicity* m_j, *then there exists a set of* $(1\times p)$ *vectors* $\{u_j^\kappa\}$ *and positive real numbers* $G_{n_j}^\kappa$ *for* $\kappa=1,2,\cdots,m_j$ *and* $j=1,2,\cdots,l$ *such that* (3.34) *holds for* $k=0,1,2,\cdots,(2n-1)$.

COROLLARY 3.3. *If* $H_n'(T)$ *is negative* (*positive*) *definite for all* n, *then the zeros of* $|P_n'(s)|$ *are also negative* (*positive*).

Proof. Theorem 3.3 can be used to write the following:

(3.37)

$$H_n'(T)=(U_nV_n)^t\left[\operatorname{diag}\left(G_{n_1}^1\alpha_1\ G_{n_1}^2\alpha_1\ \cdots\ G_{n_1}^{m_1}\alpha_1\ \cdots\ G_{n_l}^1\alpha_l\ G_{n_l}^2\alpha_l\ \cdots\ G_{n_l}^{m_l}\alpha_l\right)\right]U_nV_n.$$

Negativity (positivity) of α_j, $j=1,2,\cdots,l$ then readily follows from (3.37) by noting the facts that (U_nV_n) is of full rank and $G_{n_j}^\kappa$, $\kappa=1,2,\cdots,m_j$, $j=1,2,\cdots,l$, are all positive.

It has been demonstrated in the above how the matrix version of the Gauss quadrature formula can be used to prove the negativeness of the zeros of $|P_n'(s)|$. This property is crucial to the use of the matrix Cauchy index for demonstrating the RC-ideal transformer realizability of certain Padé approximants to any specified matrix Stieltjes series, as discussed in the following section.

4. Proof of RC realizability of the matrix Padé-approximants via the matrix Cauchy index. Recently the concept of Cauchy index of a rational function has been extended to rational matrices by Bitmead and Anderson [4]. Furthermore, criteria for rational matrices to be realizable multiport impedances have also been given in terms of the concepts of Cauchy index and McMillan degree. The purpose of this section is to show alternate ways of viewing Theorem 2.1 in the light of matrix Cauchy index and the results on matrix orthogonal polynomials dealt with in the last section. Facts from [4] will be freely made use of.

For the sake of logical development of the results it is necessary to show that a Padé approximant of appropriate order to a matrix Stieltjes series is a symmetric rational matrix. It is obvious from positive definiteness of $H_n(T)$ and from (3.1) that right matrix Padé approximant $Q_L(s)P_M^{-1}(s)$ of order $[L/M]$ exists for a matrix Stieltjes series. Since (4.1) holds and $T(s)$ is symmetric, $[Q_L(s)P_M^{-1}(s)]^t$

$$Q_L(s)P_M^{-1}(s)-T(s)=O(s^{L+M+1}) \tag{4.1}$$

is definitely a left Padé approximant of the same order. Since both right and left matrix Padé approximants exist, they must be identically equal [1, Chap. 17] and therefore, the symmetry of the approximant follows.

The following result from [4] gives a criterion for RC or RL multiport realizablity of a rational matrix in terms of matrix Cauchy index and McMillan degree. Consider a real rational $(p\times p)$ symmetric matrix $Z(s)$ and denote its matrix Cauchy index between $[a,b]$ and McMillan degree by $\mathrm{I}_a^b Z(s)$ and $\delta[Z(s)]$ respectively. Then, the following holds.

THEOREM 4.1 [4]. (i) $Z(s)$ *is realizable as an impedance of a p-port* RC *network, possibly including transformers if and only if* $I_{-\infty}^{\varepsilon} Z(s)=\delta[Z(s)]$ *for all* $\varepsilon>0$. (ii) $Z(s)$ *is realizable as an impedance of a p-port* RL *network, possibly including ideal transformers if and only if* (a) *any pole of entries of* $Z(s)$ *is of order at most one and the associated residue matrix* Z_∞ (*for a possible pole at* $s=\infty$) *is nonnegative definite* (*b*) $I_{-\infty}^{\varepsilon} Z(s)=-\delta[Z(s)]+\operatorname{rank} Z_\infty$ *for all* $\varepsilon>0$.

In the following it will be shown in several elementary steps that the above condition for RC realizability is satisfied by the $[(n-1)/n]$ and $[n/n]$ order Padé approximants to a matrix Stieltjes series. For brevity the $[(n-1)/n]$ case will be treated in detail and the case for $[n/n]$ approximants will be outlined only.

1. Note again that, in view of (3.2), (3.3), the sequence of "denominator" polynomial matrices $\{P_n(s)\}$ of Padé approximants of order $[(n-1)/n]$ are the sequence of inverse polynomial matices obtained from the sequence $\{P_n'(s)\}$. Since for a matrix Stieltjes series, $H_n'(T)$ is negative definite for all n, Corollary 3.3 implies that zeros of $|P_n'(s)|$ are negative, which in turn ensures the negativity of poles of each entry of $Q_{n-1}(s)P_n^{-1}(s)$. The eigenvalues of $Z(s)$, therefore, cannot go to infinity for any nonnegative value of s. Recalling from [4] that $\mathrm{I}_a^b Z(s)$ is the number of eigenvalues of $Z(s)$ which jump from $-\infty$ to $+\infty$ minus the number which jump from $+\infty$ to $-\infty$ as s traverses the real axis from a to b, it can be readily seen that $\mathrm{I}_{-\infty}^{+\infty}[Q_{n-1}(s)P_n^{-1}(s)]=\mathrm{I}_{-\infty}^{\varepsilon}[Q_{n-1}(s)P_n^{-1}(s)]$ for all $\varepsilon>0$.

2. Consider, next the power series expansion

$$Q_{n-1}(s)P_n^{-1}(s)=\sum_{k=0}^{\infty} R_k\cdot s^k,$$

where by virtue of the fact that $Q_{n-1}(s)P_n^{-1}(s)$ is the $[(n-1)/n]$ order Padé approximant to $T(s)=\Sigma_{k=0}^{\infty} T_k s^k$, one has $R_k=T_k$ for $k=0,1,2,\cdots,(2n-1)$. Then

$$\frac{1}{s}Q_{n-1}\left(\frac{1}{s}\right)P_n^{-1}\left(\frac{1}{s}\right)=\sum_{k=0}^{\infty} R_k\left(\frac{1}{s}\right)^{k+1}. \tag{4.2}$$

If one considers the infinite Hankel matrix associated with the power series expansion (4.2), then it is clearly seen from (4.3) that its $(np\times np)$ leading principal submatrix is $H_{n-1}(T)$, which is positive definite. Considering the fact

$$\left[\begin{array}{ccc:ccc} R_0 & \cdots & R_{n-1} & R_n & \cdots \\ R_1 & & & \cdot & \\ \vdots & & \vdots & \vdots & \\ R_{n-1} & \cdots & R_{2n-2} & R_{2n-1} & \cdots \\ \hdashline R_n & \cdots & R_{2n-1} & R_{2n} & \cdots \\ \vdots & & & \vdots & \cdots \end{array}\right] = \left[\begin{array}{ccc:cc} & & & R_n & \cdots \\ & H_{n-1}(T) & & \vdots & \\ & & & R_{2n-1} & \cdots \\ \hdashline R_n & \cdots & R_{2n-1} & R_{2n} & \cdots \\ \vdots & & & \vdots & \cdots \end{array}\right] \tag{4.3}$$

that the McMillan degree of $\frac{1}{s}Q_{n-1}(\frac{1}{s})P_n^{-1}(\frac{1}{s})$ is the rank of the infinite Hankel matrix in (4.3), it is possible to assert that this rank is at least as large as np. However, the McMillan degree cannot be larger than np. Therefore,

$$\delta\left[\frac{1}{s}Q_{n-1}\left(\frac{1}{s}\right)P_n^{-1}\left(\frac{1}{s}\right)\right]=np. \tag{4.4}$$

3. Finally, it is noted from [4] that $\mathrm{I}_{-\infty}^{+\infty}[\frac{1}{s}Q_{n-1}(\frac{1}{s})P_n^{-1}(\frac{1}{s})]$ is equal to the finite signature of the infinite Hankel matrix in (4.3), which is the signature of $H_{n-1}(T)$. Since $H_{n-1}(T)$ is positive definite symmetric, its eigenvalues, np in number, are all positive, implying that $\mathrm{I}_{-\infty}^{+\infty}[\frac{1}{s}Q_{n-1}(\frac{1}{s})P_n^{-1}(\frac{1}{s})]=np$.

It is therefore seen from the above discussion that condition (i) of Theorem 4.1 is satisfied by $\frac{1}{s}Q_{n-1}(\frac{1}{s})P_n^{-1}(\frac{1}{s})$. RC impedance realizability of $Q_{n-1}(s)P_n^{-1}(s)$ therefore follows by noting the fact that $Z(s)$ is RC impedance realizable if $\frac{1}{s}Z(\frac{1}{s})$ is also RC impedance realizable.

An exactly similar proof can be worked out for the $[n/n]$ approximant $Q_n(s)P_n^{-1}(s)$, by invoking part (ii) of Theorem 4.1 on $Q_n(\frac{1}{s})P_n^{-1}(\frac{1}{s})=\Sigma_{k=0}^{\infty}R'_k(\frac{1}{s})^k$. The required result then follows by noting that for the case under consideration $Q_n(\frac{1}{s})P_n^{-1}(\frac{1}{s})$ cannot have any pole at infinity, i.e., $Z_\infty=0$ and $\mathrm{I}_{-\infty}^{\varepsilon}Q_n(\frac{1}{s})P_n^{-1}(\frac{1}{s})=-np$. Since $Q_n(\frac{1}{s})P_n^{-1}(\frac{1}{s})$, viewed as an impedance matrix, is RL-ideal transformer realizable, $Q_n(s)P_n^{-1}(s)$ must be realizable as an impedance matrix of a RC-ideal transformer multiport.

5. Discussion and conclusion. Some network theoretic interpretations of the results of §3 will be given here. First, it is noted that since it is known from §2 that $[(n-1)/n]$ Padé approximants, when $H_n(T)$ is positive definite and $H'_n(T)$ is negative definite, are indeed impedance matrices of multiport RC networks, one could, in this special case, very well expect that the zeros of $|P_n(s)|$ would turn out to be not only real but also negative. Furthermore, in the special case, when $H_n(T)$ and $-H'_n(T)$ are positive definite, one could, via a network argument prove the Gauss quadrature formula of Theorem 3.3. This is done in the following:

Since $Q_{n-1}(s)P_n^{-1}(s)$ is a realizable RC multiport impedance, it can be written as a partial fraction expansion as:

$$Q_{n-1}(s)P_n^{-1}(s)=\sum_{j=1}^{l}\sum_{\kappa=1}^{m_j}\frac{1}{s+\beta_j}A_{n_j}^{\kappa}, \tag{5.1}$$

where β_j's are positive and $A_{n_j}^{\kappa}$'s are nonnegative definite symmetric matrices. Note that since $P_n(s)$ in (3.2) and $P'_n(s)$ in (3.3) are mutually inverse polynomial matrices, $\beta_j=-(1/\alpha_j)$.

Expanding (5.1),

$$\begin{aligned}Q_{n-1}(s)P_n^{-1}(s)&=\sum_{j=1}^{l}\sum_{\kappa=1}^{m_j}\left(\sum_{k=0}^{\infty}s^k\alpha_j^k\right)\left(\frac{1}{\beta_j}A_{n_j}^{\kappa}\right)\\&=\sum_{k=0}^{\infty}\left(\sum_{j=1}^{l}\sum_{\kappa=1}^{m_j}\left(\frac{1}{\beta_j}A_{n_j}^{\kappa}\right)\alpha_j^{\kappa}\right)s^k.\end{aligned} \tag{5.2}$$

Now because of the fact that $Q_{n-1}(s)P_n^{-1}(s)$ is a Padé approximant of order $[(n-1)/n]$ to $T(s)=\Sigma_{k=0}^{\infty}T_ks^k$, comparison of coefficients of s^k, $k=0,1,2,\cdots,(2n-1)$, in the power series expansion for $T(s)$ and in (5.2), one obtains,

$$T_k=\sum_{j=1}^{l}\sum_{\kappa=1}^{m_j}\left(\frac{1}{\beta_j}A_{n_j}^{\kappa}\right)\alpha_j^k \tag{5.3}$$

for $k=0,1,2,\cdots,(2n-1)$.

It is noted that since $(\frac{1}{\beta_j}A_{n_j}^{\kappa})$ is a nonnegative definite matrix, (5.3) is nothing but the Gauss quadrature formula in different form. Interestingly, this formulation also establishes a relation between the residue matrices $A_{n_j}^{\kappa}$'s and $K_n(\alpha_j,\alpha_j)$, the kernel polynomial evaluated at α_j, via the use of Theorem 3.3.

The primary objective of the paper, therefore, has been the demonstration of the relationship between rational approximants of appropriate orders to a specified symmetric matrix power series of a special type and multiport network synthesis using the matrix version of the classical continued fraction expansion theory and the recently developed artifice of matrix Cauchy index. The 'denominator' polynomial matrices of the approximants are shown to form an orthogonal polynomial matrix sequence over a real semi-infinite interval. Though mathematical derivations of the properties of the

polynomial matrices belonging to the orthogonal polynomial matrix sequence have been given, the reader has also been alerted to the feasibility of network-theoretic justification of the results. It is hoped that these links between mathematical results and the theory of network realizability will kindle interest for further research among scientists coming from either discipline.

Acknowledgment. The authors wish to extend their sincere thanks to the reviewer for helpful and constructive criticisms of the paper.

REFERENCES

[1] G. A. BAKER, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] KUCHIRO MORIMOTO, NAOKI MATSUMOTO AND SHIN-ICHI TAKAHASHI, *Matrix Padé approximants and multiport networks*, Electronics and Communications in Japan, 61A (1978), pp. 28–36.

[3] LEANG-SAN SHIEH AND FRANK F. GAUDIANO, *Some properties and applications of matrix continued fractions*, IEEE Trans. Circuits and Systems, CAS-32 (1975), pp. 721–728.

[4] R. BITMEAD AND B. D. O. ANDERSON, *The matrix Cauchy index: properties and applications*, SIAM J. Appl. Math., 33 (1977), pp. 655–672.

[5] PH. DELSARTE, Y. V. GENIN AND Y. V. KAMP, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 149–160.

[6] H. S. WALL, *Analytic Theory of Continued Fractions*, Chelsea, New York, 1948.

[7] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[8] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

[9] W. B. BRAGG, *Matrix interpretation and applications of continued fraction algorithm*, Rocky Mountain J. Math., 4 (1974), pp. 213–225.

[10] D. C. YOULA AND N. KAZANJIAN, *Bauer type factorization of positive matrices and theory of matrix orthogonal polynomials on the unit circle*, IEEE Trans. Circuits and Systems, CAS 25 (1978), pp. 57–69.

[11] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0002 $01.25/0

MEAN VALUE AND TAYLOR FORMS IN INTERVAL ANALYSIS*

L. B. RALL†

Abstract. Basic spaces for interval analysis are constructed as Cartesian products of the real line. The spaces obtained in this way include real finite- and infinite-dimensional real vector spaces, and have a number of important Hilbert and Banach spaces as subspaces in the sense of set inclusion. A Gâteaux-type derivative is defined in these spaces, and is used in the corresponding interval spaces, together with interval arithmetic, to obtain interval versions of the mean value theorem and Taylor's theorem. These theorems provide ways to construct accurate interval inclusions of operators, called mean value and Taylor forms. The forms resulting from expansion about midpoints of intervals are shown to be inclusion monotone, and the effect of outward rounding on this class of forms is also considered. An application is made to show that interval iteration operators for the solution of operator equations can be constructed which have arbitrarily high order of convergence in width. Derivations of the fundamental theorems of less generality from results in real and functional analysis are also presented. As in the case of real and functional analysis, the interval Taylor's theorem given here provides a powerful tool for applications of interval analysis to problems in applied mathematics.

Key words. interval analysis, calculus in abstract spaces, mean value theorem, Taylor's theorem, interval inclusions, interval iteration, fixed point problems, solution of equations

1. A setting for interval analysis. In the same way that real analysis is concerned with transformations of real numbers (or vectors) into others, *interval analysis* [6], [7] deals with transformations of intervals (or interval vectors). Since an ordering relationship is fundamental to the definition of intervals, a natural abstract setting for interval analysis is a partially ordered space [1], [14], or, more specifically, a lattice [1], [8]. Here, a more concrete approach will be taken, which results in the construction of what will be called *IR-spaces* by forming Cartesian products of the set IR of nonempty closed intervals

$$X=[a,b]=\{x|a\leq x\leq b,\ x\in R\}, \tag{1.1}$$

on the real line R. Interval analysis on these IR-spaces will be called *real* interval analysis; it is general enough to cover many important applications, and the theory obtained adapts readily to actual numerical computation, for which only a finite set of real numbers is available.

1.1. Real spaces. The spaces to be considered here are built in a natural way from the set R of real numbers. Given a set A, one can form the *Cartesian product*

$$P=\prod_A R \tag{1.2}$$

of R over the *index set* A to obtain a set of *vectors* f with real *components* $f_\alpha\in R$, $\alpha\in A$. Writing $f=\{f_\alpha|\alpha\in A,\ f_\alpha\in R\}$ for $f\in P$, P is a *linear* space for the *componentwise* definitions of *addition* $f+g$ and multiplication by *scalars* (real numbers) $a\cdot f$ given by

$$f+g=\{f_\alpha+g_\alpha|\alpha\in A\},\qquad a\cdot f=\{a\cdot f_\alpha|\alpha\in A\}, \tag{1.3}$$

respectively [12], [16].

*Received by the editors September 30, 1981. This research was sponsored by the U. S. Army under contract DAAG29-80-C-0041.

†Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

Another way of looking at the product space P given by (1.2) is as the set of all *functionals* (real-valued functions) f on A; one writes $f_\alpha = f(\alpha)$, and (1.3) gives the natural definitions of sums and scalar multiples of functionals, which are also functionals.

DEFINITION 1.1. A *real space* (or *R-space* for short) is a linear space P constructed according to (1.2) and (1.3) or the Cartesian product

$$P = \prod_{\beta \in B} P_\beta \tag{1.4}$$

of such spaces, again with addition and multiplication by scalars defined componentwise.

Examples of R-spaces abound. The choice $A = \{1, 2, \cdots, n\}$ in (1.2) gives $P = R^n$, the space of n-dimensional real vectors $f = (f_1, f_2, \cdots, f_n)$, while $A = \{1, 2, 3, \cdots\}$, the set of positive integers, gives R^∞, which consists of the real sequential vectors $f = (f_1, f_2, f_3, \cdots)$. Going on to $A = X = [a, b]$, a nonempty interval (1.1), one gets $P = R[a, b]$, the space of all real functions f on the interval $[a, b]$, the components of which are usually denoted by $f(x) = f_x$, $a \le x \le b$. Similarly, if $Y = [c, d]$ is also an interval, then taking $A = X \times Y = [a, b] \times [c, d]$ gives the space $R([a, b] \times [c, d])$ of real functions f of two variables with components $f(x, y)$, $a \le x \le b$, $c \le y \le d$, and so on.

Cartesian products (1.4) can be used for concise description of sets of functions taking on values in R-spaces. For example, with $X = [a, b]$, $Y = [c, d]$, the real space $R(X \times Y) \times R(X \times Y)$ consists of all functions $f: X \times Y \subset R^2 \to R^2$ with components $f(x, y) = (f_1(x, y), f_2(x, y))$, $a \le x \le b$, $c \le y \le d$. More generally, if $D \subset P$ and Q is a real space, then

$$\prod_D Q = \{ f \mid f: D \subset P \to Q \} \tag{1.5}$$

is also a real space by Definition 1.1. In (1.5), it is not required that P be a real space, but this will usually be the case in the following discussion. A simple, but important, example of (1.5) is obtained for $P = R^n$, $Q = R^m$, which gives the set of functions (or *operators*) $f: D \subset R^n \to R^m$, which are fundamental to computational numerical analysis [10], [12].

The subject of *functional analysis* is concerned with analysis on *normed* linear spaces (usually the ones which are *complete*, called *Banach* spaces) [12], [16]. A number of useful spaces of this type over the real scalar field can be considered to be subspaces of real spaces P in the sense that all their elements belong to P. In particular, all finite-dimensional real normed linear spaces are pretty much indistinguishable, due to the equivalence of norms [16], and can be identified with the real spaces R^n. The situation is different for infinite-dimensional spaces. For example, the elements of the Banach space R_∞^∞ of sequential real vectors f such that

$$\|f\|_\infty = \sup_{(n)} \{ |f_n| \} < +\infty, \tag{1.6}$$

form a subspace of R^∞ which is different from the one consisting of elements of R_2^∞, for which

$$\|f\|_2 = \left\{ \sum_{n=1}^{\infty} f_n^2 \right\}^{1/2} < +\infty. \tag{1.7}$$

Similarly, the space $C[a,b]$ of *continuous* functions f on $a \le x \le b$ (with the usual norm) can be identified with a subspace of $R[a,b]$ which is different from the one obtained from $L_2[a,b]$, which consists of $f \in R[a,b]$ such that

$$\|f\|_2 = \left\{ (L) \int_a^b f(x)^2 dx \right\}^{1/2} < +\infty, \tag{1.8}$$

where (L) denotes Lebesgue integration [16]. This natural type of embedding of normed linear spaces into real spaces will be helpful below in connection with the derivation of interval versions of results from real and functional analysis.

1.2. Real interval spaces. The set of finite, nonempty intervals (1.1) on the real line R will be denoted by IR. There is a natural identification of real numbers $x \in R$ with *degenerate* intervals $[x,x] \in IR$ with equal endpoints, and one writes

$$x = [x,x]. \tag{1.9}$$

Ordinary arithmetic, extended from R to IR, is called *interval arithmetic* [6], [7]. For example, *addition* of intervals $X=[a,b]$ and $Y=[c,d]$ is defined by

$$X+Y=[a,b]+[c,d]=[a+c,b+d], \tag{1.10}$$

and *multiplication* of $X=[a,b]$ by a real number $r=[r,r]$ by

$$r \cdot X = \begin{cases} [ra,rb], & r \ge 0, \\ [rb,ra], & r < 0. \end{cases} \tag{1.11}$$

Note that with these definitions, IR is *not* a linear space; with subtraction defined in the usual way by $X-Y=X+(-1)\cdot Y$, (1.10) and (1.11) give

$$[0,1]-[0,1]=[-1,1] \tag{1.12}$$

instead of the identity element $0=[0,0]$ of interval addition.

It will be useful to associate the following real numbers with an interval $X=[a,b] \in IR$: Its *midpoint* $m[a,b]=(a+b)/2$, its *width* $w(X)=w[a,b]=b-a$, and its *absolute value* (or *modulus*) $|X|=|[a,b]|=\max\{|a|,|b|\}$ [7].

Another important property of intervals is that the *intersection* $X \cap Y=[a,b] \cap [c,d]$ is either the interval

$$X \cap Y=[a,b] \cap [c,d]=[\max\{a,c\}, \min\{b,d\}] \tag{1.13}$$

or the *empty set* $\varnothing$; if $b<c$ or $a>d$, then

$$X \cap Y = \varnothing; \tag{1.14}$$

otherwise, (1.13) holds. Furthermore, if $\{X_n\}$ is a sequence of *nested* intervals, that is

$$X_1 \supset X_2 \supset X_3 \supset \cdots, \tag{1.15}$$

then

$$X = \bigcap_{n=1}^{\infty} X_n \neq \varnothing, \tag{1.16}$$

since each X_n is a closed, nonempty subset of R [15].

The construction (1.2), (1.4) of real spaces in §1.1 will now be used to obtain the corresponding interval spaces, by starting with IR in place of R.

DEFINITION 1.2. A *real interval space* IP is a space of the form

$$IP = \prod_A IR, \tag{1.17}$$

or

$$IP=\prod_{\beta\in B} IP_\beta, \tag{1.18}$$

in which each real interval space IP_β is of the form (1.17). Real interval spaces will also be referred to as *IR-spaces*.

There is an obvious one-to-one correspondence between interval spaces (1.17), (1.18) and real spaces (1.2), (1.4), respectively. Furthermore, the order relationships $<,\leq,\geq,>$ in R can be extended componentwise to a real space P to obtain a *partial ordering* [1] of P. In the resulting partial ordering, the corresponding IR-space IP consists of the set of all *intervals* in P; that is, $X\in IP$ if and only if there are elements $a,b\in P$ such that $a\leq b$ and $X=[a,b]=\{x\,|\,a\leq x\leq b,\ x\in P\}$, which is (1.1) with R replaced by P. This leads to the embedding $x=[x,x]$ of P into IP, as in (1.9). Moreover, interval arithmetic is also extended componentwise from IR to an arbitrary real interval space IP. As in the case of IR, IP will not be a linear space, unlike its underlying R-space P. The quantities $m(X)$, $w(X)$, and $|X|$ defined previously for real intervals $X\in IR$ can also be defined componentwise for $X\in IP$, with the result being that $m(X)$, $w(X)$, and $|X|$ will be elements of P.

Typical examples of IR-spaces are the space IR^n of *interval vectors*

$$X=(X_1,X_2,\cdots,X_n),\qquad X_i\in IR,\qquad i=1,2,\cdots,n, \tag{1.19}$$

and the space $IR[a,b]$ of *interval functions* Y on $[a,b]\in IR$ defined by

$$Y(x)=[c(x),d(x)],\qquad a\leq x\leq b, \tag{1.20}$$

where $c,d\in R[a,b]$ and $c\leq d$ [3], [13]. For $X\in IR^n$, for example, one has

$$m(X)=(m(X_1),m(X_2),\cdots,m(X_n))\in R^n, \tag{1.21}$$

and for $Y\in IR[a,b]$, $|Y|$ is defined by

$$|Y|(x)=|Y(x)|=\max\{|c(x)|,|d(x)|\},\qquad a\leq x\leq b, \tag{1.22}$$

and thus $|Y|\in R[a,b]$ is a real function.

An interval $X\in IP$ is, by construction, a subset of the underlying R-space P. One important property of intervals in IP as subsets of P is *convexity*.

LEMMA 1.1. *If P is a real space and $X\in IP$, then X is a convex subset of P; that is, for arbitrary points $x,y\in X$,*

$$\Lambda(x,y)=\{z\,|\,z=\theta y+(1-\theta)x,\ 0\leq\theta\leq 1\}\subset X. \tag{1.23}$$

Proof. It follows from Definition 1.1 that each $f\in P$, P a real space, can be represented as $f=\{f_\gamma\,|\,f_\gamma\in R,\ \gamma\in B\times A=\Gamma\}$, the real numbers $f_\gamma,\gamma\in\Gamma$, being the components of f. Now, let $X=[a,b]$, and define $c,d\in P$ by

$$c_\gamma=\min\{x_\gamma,y_\gamma\},\qquad d_\gamma=\max\{x_\gamma,y_\gamma\},\qquad \gamma\in\Gamma. \tag{1.24}$$

For $x,y\in X$, it follows that

$$a_\gamma\leq c_\gamma\leq\theta y_\gamma+(1-\theta)x_\gamma\leq d_\gamma\leq g_\gamma,\qquad \gamma\in\Gamma, \tag{1.25}$$

for $0\leq\theta\leq 1$; hence, from (1.23), $\Lambda(x,y)\subset X$. Q.E.D.

As usual, the set $\Lambda(x,y)$ defined by (1.23) is called the *line segment from x ($\theta=0$) to y ($\theta=1$)*. A useful class of intervals is the symmetric intervals, defined as follows:

DEFINITION 1.3. An interval $S\in IP$ is said to be *symmetric* if $-s\in S$ for each $s\in S$.

As a consequence of this definition, each symmetric interval S contains the *origin* 0 of P; furthermore, $S=[-a,a]$ for some element $a\geq 0$ of P. Moreover, if $s\in P$, then

$S=[-1,1]\cdot s$ will be a symmetric interval; in this case, one can write $S=[-1,1]\cdot s=s\cdot[-1,1]=[-|s|,|s|]$, where $|s|$ is the absolute value of s defined componentwise in the usual way.

In addition to the origin 0 of a real space P (the element such that $0_\gamma=0$, $\gamma\in\Gamma$), it is helpful to single out the element $e\in P$ defined by $e_\gamma=1$, $\gamma\in\Gamma$. In terms of e, the symmetric intervals $S_\rho\in IP$ are defined for real $\rho\geq 0$ by

$$S_\rho=\rho\cdot[-e,e]=\rho e\cdot[-1,1],\qquad \rho\in R,\quad \rho\geq 0. \tag{1.26}$$

DEFINITION 1.4. A set $D\subset P$ is said to be *bounded* if

$$D\subset S_\rho=\rho e\cdot[-1,1] \tag{1.27}$$

for some real ρ such that $0\leq\rho<+\infty$, in particular, if D consists of a single element $f\in P$, then f is called a *bounded element* of P.

2. Interval transformations. Suppose that IP, IQ are IR-spaces, and $F: ID\subset IP\to IQ$ is an operator defined on a domain ID in IP which takes on values in IQ. The result of applying F to $X\in ID$ is symbolized by

$$Y=F(X), \tag{2.1}$$

where $Y\in IQ$, and F is called an *interval transformation* from $ID\subset IP$ into IQ. It follows that $F\in\Pi_{ID}IQ$. What will be called *interval analysis* here refers to the study of interval transformations.

DEFINITION 2.1. The interval transformation $F: ID\subset IP\to IQ$, where IP, IQ are real interval spaces, is said to have an *interval domain* ID if $Z\in ID$ implies that $X\in ID$ for each subinterval $X\subset Z$ of Z.

An important class of interval transformations are the ones which are monotone in the sense of the following definition.

DEFINITION 2.2. An interval transformation $F: ID\subset IP\to IQ$ with interval domain ID is said to be *inclusion monotone* (or simply *monotone*) on ID if

$$X\subset Z\Rightarrow F(X)\subset F(Z) \tag{2.2}$$

for each $Z\in ID$.

Given a domain $D\subset P$, P a real space, the corresponding interval domain ID in IP can be constructed from the set of intervals $Z\subset D$ (which includes all the degenerate intervals equivalent to points of D) by adjoining all subintervals of each such Z, if necessary. In what follows, it will be assumed that domains ID for interval transformations corresponding to domains D of real transformations are formed in this way, and hence will be interval domains. One has also $D\subset ID$ by the identification of points of P with degenerate intervals in IP. The concept of an interval domain corresponding to a real one leads to a fundamental relationship between real and interval transformations.

DEFINITION 2.3. The interval transformation $F: ID\subset IP\to IQ$ is said to be an *inclusion* of the real transformation $f: D\subset P\to Q$ between the underlying real spaces P and Q if

$$f(X)=\{f(x)\,|\,x\in X\}\subset F(X) \tag{2.3}$$

for each $X\subset D$. If F is monotone on ID, then it is called a *monotone inclusion* of f.

For most of the results to be obtained below, inclusions of real transformations are adequate. However, the property of monotonicity is highly desirable in many applications. Some interval inclusions of real transformations also have the following property.

DEFINITION 2.4. The interval inclusion $F: ID \subset IP \to IQ$ of $f: D \subset P \to Q$ is said to have the *restriction property* on D if

$$F(x) = F([x,x]) = f(x) \tag{2.4}$$

for each $x \in D$, in which case F is called an *interval extension* of f on D. If F is also monotone on D, then it is called a *monotone interval extension* of f on D.

The rules of interval arithmetic [6], [7] are examples of monotone interval extensions, in this case of the real transformations $f: R^2 \to R$ defined by $f(x,y) = x \circ y$ for $\circ = +, -, \cdot, /$. (For division, $D = R \times (R \setminus \{0\})$, of course.) In actual computation, one ordinarily has to forego the restriction property (2.4), since it is impossible to represent arbitrary real numbers exactly with the finite set of numbers available on a given computer. The use of interval arithmetic and *directed* (or perhaps *outward*) *rounding*, however, allows one to construct monotone inclusions of rational functions automatically, even if the endpoints of intervals have to be selected from a finite set of numbers G, provided that the computation stays within the interval $IG = [\min\{G\}, \max\{G\}]$ [6], [7]. Along with interval arithmetic, there are other methods for the construction of interval inclusions of real transformations. The ones to be discussed in this paper are based on interval versions of the mean value theorem and Taylor's theorem in ordinary real analysis [4].

3. A derivative in *R*-spaces. As usual, if P is a real space, then a function $f: D \subset R \to P$ will be called an *abstract function*; for example, $z: R \to P$ defined for $x, y \in P$ by

$$z(\theta) = \theta y + (1-\theta)x = x + \theta(y-x), \qquad \theta \in R, \tag{3.1}$$

takes on values on the *line* through x, y for $x \neq y$ (see (1.23)). For $f: D \subset R \to P$, where D contains a neighborhood of 0, it is said that

$$\lim_{\theta \to 0} f(\theta) = 0 \tag{3.2}$$

if there is a real-valued function $\rho \geq 0$, monotone decreasing in $|\theta|$, such that

$$f(\theta) \in \rho(\theta) e \cdot [-1,1] \quad \text{and} \quad \lim_{\theta \to 0} \rho(\theta) = 0. \tag{3.3}$$

DEFINITION 3.1. A function $f: D \subset P \to Q$, D convex, is said to be *differentiable* at $x \in D$ if a linear operator, denoted by $f'_D(x)$ or simply $f'(x)$, exists from the linear space LD spanned by D into Q such that

$$\lim_{\theta \to 0} \frac{r_{x,y}(\theta)}{\theta} = 0, \tag{3.4}$$

where

$$r_{x,y}(\theta) = f(x + \theta(y-x)) - f(x) - f'(x) \cdot \theta(y-x). \tag{3.5}$$

The operator $f'(x)$, easily seen to be unique if it exists, is of course called the *derivative* of f at $x \in D$. (The linear space LD referred to in Definition 3.1 is simply the set of all linear combinations of elements of D [16].) Defined in this way, f' is a derivative of Gâteaux type; in fact, if D is a Banach subspace of P such that condition (3.4) and $\lim_{\theta \to 0} \| \frac{1}{\theta} \cdot r_{x,y}(\theta) \| = 0$ are equivalent (such as R^n, R^∞_∞, and $C[0,1]$), then $f'(x)$ is precisely the Gâteaux derivative in D of f at x [4], [10]. Because of the dependence of $f'(x) = f'_D(x)$ on the domain D, this derivative can also be considered to be a type of *directional* derivative; for example, one can take D to be the line through the origin of P

consisting of the points defined by (3.1). In case it is desirable to distinguish the derivative defined above from some other derivative, it will be called the *elementary* real derivative, or simply the *R-derivative* of f at $x \in D$.

4. Elementary mean value forms.

THEOREM 4.1. *If X is an interval such that f is differentiable on $X \cap D$, D convex, and F' is an interval inclusion of f' on X, then*

$$f(y)-f(x) \in F'(X)\cdot(X-x), \qquad x,y \in X \cap D. \tag{4.1}$$

Proof. For $x,y \in X \cap D$, it follows from Definition 3.1 that given a real $\varepsilon>0$, there exists a real number τ, $0<\tau\leq 1$, such that

$$f(x+\theta(y-x))-f(x) \in F'(X)\cdot\theta(y-x)+\varepsilon\theta e\cdot[-1,1] \tag{4.2}$$

for $0\leq\theta\leq\tau$. To show that (4.2) holds for $\theta=1$, the assumption that $\tau<1$ is the supremum of the values for which it is valid will now be contradicted. Set $z=x+\tau(y-x)$. Since $f'(z)$ exists, there is a real number β, $\tau<\beta\leq 1$ such that

$$f(x+\eta(y-x))-f(z) \in F'(X)\cdot(\eta-\tau)(y-x)+\varepsilon(\eta-\tau)e\cdot[-1,1], \tag{4.3}$$

$\tau\leq\eta\leq\beta$. Let $\theta=\tau$ in (4.2) and add to (4.3) to obtain

$$f(x+\eta(y-x))-f(x) \in F'(X)\cdot\eta(y-x)+\varepsilon\eta e\cdot[-1,1], \tag{4.4}$$

$\tau\leq\eta\leq\beta$, and thus (4.2) holds for $0\leq\theta\leq\beta$, which contradicts the assumed property of τ, since $\beta>\tau$. Hence, for $\theta=1$, $\varepsilon=\frac{1}{n}$, (4.2) becomes

$$f(y)-f(x) \in F'(X)(X-x)+\frac{e}{n}\cdot[-1,1]. \tag{4.5}$$

It follows that

$$f(y)-f(x) \in \bigcap_{n=1}^{\infty}\left\{F'(X)\cdot(X-x)+\frac{e}{n}\cdot[-1,1]\right\}=F'(X)\cdot(X-x)+[0,0], \tag{4.6}$$

which is nothing more nor less than (4.1). Q.E.D.

The proof of Theorem 4.1 given above is truly elementary in that only interval arithmetic and the definitions of interval inclusions and the derivative are used. Replacing X by $\Lambda(x,y)$ in the above proof leads to the conclusion

$$f(y)-f(x) \in F'(\Lambda(x,y))\cdot(y-x) \subset F'(\Lambda(x,y))\cdot(X-x), \tag{4.7}$$

which is also valid. If F' is a monotone inclusion of f', then (4.7) implies (4.1). Note that f' need not be defined on all of X; all that is required is that $f'(X\cap D)\subset F'(X)$; one can take $f'(x)=F'([x,x])$ for $x \in X\setminus\{X\cap D\}$.

DEFINITION 4.1. If F' is an interval inclusion of f' on X, then the interval inclusion F of f on X defined by

$$F(X)=f(x)+F'(X)\cdot(X-x) \tag{4.8}$$

is called the (*elementary*) *mean value form* of f.

The mean value form was introduced by Moore [6] in R^n, and studied in R^n and $C'[a,b]$ by Caprani and Madsen [2], whose results will be returned to below. The form (4.8) provides a method, in addition to interval arithmetic, for the construction of interval inclusions of real transformations. A useful case of the mean value form is its *midpoint* (or *centered*) form, obtained for $x=m(X)$. Since

$$X=m(X)+\tfrac{1}{2}w(x)\cdot[-1,1] \tag{4.9}$$

for an arbitrary interval X and $w(X)\geq 0$, one has, if $m(X)\in D$,

$$F(X)=f(m(X))+\tfrac{1}{2}|F'(X)|w(X)\cdot[-1,1] \tag{4.10}$$

in this case, which expresses $F(X)$ as the sum of the point $f(m(X))\in Q$ and a symmetric interval in IQ. The midpoint form (4.10) is even simpler in case $y\in D$ and

$$X=X(y,\rho)=y+\rho e\cdot[-1,1] \tag{4.11}$$

is the *cube* with center $y\in P$ and radius ρ. Then,

$$F(X(y,\rho))=f(y)+\rho|F'(X(y,\rho))|e\cdot[-1,1], \tag{4.12}$$

which often can be computed very economically.

The following theorem, which is a generalization of the fundamental result due to Caprani and Madsen [2], shows that F defined by the midpoint mean value form (4.10) is monotone.

THEOREM 4.2. *If F' is a monotone inclusion of f', then F defined by the midpoint mean value form* (4.10) *is monotone on the set of intervals X such that $m(X)\in D$.*

Since F is already an inclusion of f on the set of intervals cited, all that needs to be established is monotonicity. The following lemma is the key to the proof.

LEMMA 4.1 (Caprani–Madsen [1]). *If X, Z are intervals in a real space P, then*

$$X\subset Z\Leftrightarrow \tfrac{1}{2}w(Z)\geq\tfrac{1}{2}w(X)+|m(Z)-m(X)|. \tag{4.13}$$

Proof. Suppose the inequality in (4.13) holds. Then, for $x\in X$,

$$x-m(Z)=x-m(X)+\{m(X)-m(Z)\}\in\{\tfrac{1}{2}w(X)+|m(Z)-m(X)|\}\cdot[-1,1], \tag{4.14}$$

so that $x\in m(Z)+\frac{1}{2}w(Z)\cdot[-1,1]=Z$, and thus $X\subset Z$. On the other hand, suppose that $X\subset Z$, or

$$m(X)+\tfrac{1}{2}w(X)\cdot[-1,1]\subset m(Z)+\tfrac{1}{2}w(Z)\cdot[-1,1]. \tag{4.15}$$

Since $w(Z)\geq w(X)\geq 0$, this gives

$$m(X)-m(Z)\in\tfrac{1}{2}\{w(Z)-w(X)\}\cdot[-1,1], \tag{4.16}$$

and the inequality in (4.13) follows from multiplication by $[-1,1]$. Q.E.D.

Proof of Theorem 4.2. Suppose that $U\subset V$, where $U, V\in IP$ are such that $m(U), m(V)\in D$. Set $X=F(U)$, $Z=F(V)$. It follows that $m(X)=f(m(U))$, $m(Z)=f(m(V))$, $\frac{1}{2}w(X)=|F'(U)|w(U)/2$, $\frac{1}{2}w(Z)=|F'(V)|w(V)/2$. Since $U\subset V$, one has $m(U), m(V)\in V$, and, from the proof of Theorem 4.1,

$$f(m(V))-f(m(U))\in F'(V)\{m(V)-m(U)\}, \tag{4.17}$$

so that

$$f(m(V))-f(m(U))\in|F'(V)||m(V)-m(U)|\cdot[-1,1]. \tag{4.18}$$

For $x\in X$, $x-m(Z)=x-m(X)+f(m(U))-f(m(V))$, and

$$x-m(X)\in\tfrac{1}{2}|F'(U)|w(U)\cdot[-1,1]\subset\tfrac{1}{2}|F'(V)|w(U)\cdot[-1,1], \tag{4.19}$$

since the monotonicity of F' implies that $|F'(V)|\geq|F'(U)|$ for $U\subset V$. Using (4.18) and (4.19), one gets

$$x-m(Z)\in|F'(V)|\{\tfrac{1}{2}w(U)+|m(V)-m(U)|\}\cdot[-1,1]\subset\tfrac{1}{2}|F'(V)|w(V)\cdot[-1,1] \tag{4.20}$$

by the Caprani–Madsen Lemma 4.1, so that $x \in Z$, and thus $U \subset V \Rightarrow F(U) \subset F(V)$. Q.E.D.

Monotonicity is often crucial in numerical computation, in which only a finite set of points G and corresponding intervals IG are available. When an interval $X \in IP$ is approximated by an interval $Z \in IG \subset IP$ such that $X \subset Z$ (this process is called *outward rounding*), one wants to be sure that $F(X) \subset F(Z)$ in order for the results actually computed to contain the ones that would be obtained by exact computation.

5. Elementary Taylor forms. It can be verified without difficulty that the elementary derivative defined in §3 has the ordinary properties of a Gâteaux derivative; for example, $(f+g)' = f' + g'$ and the chain rule holds; proofs will be omitted here. Furthermore, successive differentiations give rise to *multilinear* operators from P into Q in the usual way [4], [10], [12]. The following result is an interval version of Taylor's theorem of real analysis.

THEOREM 5.1. *If f is differentiable n times on $X \cap D$, D convex, and $F^{(n)}$ is an interval inclusion of $f^{(n)}$ on X, then for $x, y \in X \cap D$,*

$$f(y) - f(x) - \sum_{k=1}^{n-1} \frac{1}{k!} f^{(k)}(x)(y-x)^k \in \frac{1}{n!} F^{(n)}(X) \cdot (X-x)^n. \tag{5.1}$$

Proof. The proof will be carried out by mathematical induction. Theorem 4.1 shows that (5.1) is valid for $n = 1$, and it will be assumed to hold for $n = m-1$. If ϕ is an abstract function which is differentiable on $[0, 1]$, then, given any $\varepsilon > 0$, it follows as in the proof of Theorem 4.1 that there exists a finite sequence of points $\{\theta_i\}_{i=0}^{\nu}$, $0 = \theta_0 < \theta_1 < \cdots < \theta_{\nu-1} < \theta_\nu = 1$, such that

$$\phi(\theta_i) - \phi(\theta_{i-1}) \in \phi'(\theta_{i-1})(\theta_i - \theta_{i-1}) + \varepsilon(\theta_i - \theta_{i-1}) e \cdot [-1, 1]. \tag{5.2}$$

For the particular abstract function

$$\phi(\theta) = f(x + \theta(y-x)) - f(x) - \sum_{k=1}^{m-1} \frac{1}{k!} f^{(k)}(x) \theta^k (y-x)^k, \tag{5.3}$$

one has $\phi(0) = 0$, and thus

$$\phi(1) - \phi(0) = \phi(1) = f(y) - f(x) - \sum_{k=1}^{m-1} \frac{1}{k!} f^{(k)}(x)(y-x)^k, \tag{5.4}$$

and

$$\phi'(\theta) = f'(x + \theta(y-x))(y-x) - f'(x)(y-x) - \sum_{k=2}^{m-1} \frac{1}{(k-1)!} f^{(k)}(x) \theta^{k-1} (y-x)^k. \tag{5.5}$$

By the induction hypothesis,

$$\phi'(\theta) \in \frac{1}{(m-1)!} F^{(m)}(X) \cdot (X-x)^m \theta^{m-1} [0, 1]. \tag{5.6}$$

Therefore, from (5.2),

$$\phi(\theta_i) - \phi(\theta_{i-1}) \in \frac{1}{(m-1)!} F^{(m)}(X) \cdot (X-x)^m \theta_{i-1}^{m-1} (\theta_i - \theta_{i-1})[0, 1] + \varepsilon(\theta_i - \theta_{i-1}) e \cdot [-1, 1], \tag{5.7}$$

$i = 1, 2, \cdots, \nu$. Thus,

$$\phi(1)-\phi(0)=\sum_{i=1}^{\nu}\{\phi(\theta_i)-\phi(\theta_{i-1})\}$$

$$\in\frac{1}{(m-1)!}F^{(m)}(X)(X-x)^m\sum_{i=1}^{\nu}\theta_{i-1}^{m-1}(\theta_i-\theta_{i-1})\cdot[0,1]+\varepsilon e\cdot[-1,1]. \tag{5.8}$$

However,

$$0<\sum_{i=1}^{\nu}\theta_{i-1}^{m-1}(\theta_i-\theta_{i-1})<\int_0^1\theta^{m-1}d\theta=\frac{1}{m}, \tag{5.9}$$

since the sum is a lower Riemann sum for the integral. Since X is convex and $x\in X$ implies $0\in(X-x)$, it follows that $(X-x)\cdot\alpha[0,1]=\alpha(X-x)\subset(X-x)\cdot\beta[0,1]=\beta(X-x)$ for $0<\alpha<\beta$. Using this fact and

$$\bigcap_{\varepsilon\to 0}\varepsilon e\cdot[-1,1]=[0,0], \tag{5.10}$$

one has

$$\phi(1)\in\frac{1}{m!}F^{(m)}(X)\cdot(X-x)^m+[0,0]=\frac{1}{m!}F^{(m)}(X)\cdot(X-x)^m, \tag{5.11}$$

which is equivalent to (5.1) with $n=m$ by (5.4). This completes the proof of the theorem by mathematical induction. Q.E.D.

Once again, little more than interval arithmetic is required in the proof.

DEFINITION 5.1. If $f: D\subset P\to Q$ is differentiable n times on $X\cap D$, $X\in IP$, then for $x\in X\cap D$,

$$F(X)=f(x)+\sum_{k=1}^{n-1}\frac{1}{k!}f^{(k)}(x)\cdot(X-x)^k+\frac{1}{n!}F^{(n)}(X)\cdot(X-x)^n, \tag{5.12}$$

where $F^{(n)}$ is an interval inclusion of $f^{(n)}$ on X, is called the (*elementary*) *Taylor form of f of order n*.

It follows from Theorem 5.1 that F defined by (5.12) is an interval inclusion of f on X. For the particular choice $x=m(X)$, one obtains the *midpoint* form of (5.12),

$$F(X)=f(m(X))+\left\{\sum_{k=1}^{n-1}\frac{1}{2^k k!}|f^{(k)}(m(X))|w(X)^k+\frac{1}{2^n n!}|F^{(n)}(X)|w(X)^n\right\}\cdot[-1,1], \tag{5.13}$$

and, for $X=X(y,\rho)$ a cube, the *cube-centered* form

$$F(X(y,\rho))=f(y)+\left\{\sum_{k=1}^{n-1}\frac{\rho^k}{k!}|f^{(k)}(y)|e^k+\frac{\rho^n}{n!}|F^{(n)}(X(y,\rho))|e^n\right\}\cdot[-1,1]. \tag{5.14}$$

Evaluations of this latter form can often be carried out very economically, since operations on e ordinarily do not require multiplications, and only nonnegative operators are involved. Monotonicity of the midpoint Taylor form (5.13) also follows from monotonicity of $F^{(n)}$, in much the same way as for the midpoint mean value form (4.10).

THEOREM 5.2. *If $F^{(n)}$ is a monotone inclusion of $f^{(n)}$, then F defined by the midpoint Taylor form* (5.13) *is monotone on the set of intervals X such that $m(X)\in D$.*

Proof. As before, suppose $U \subset V$, and it is to be shown that $F(U) \subset F(V)$, where the results of the transformations of U, V by F are given by (5.13). For brevity of notation, set $u = m(U)$, $v = m(V)$. It follows that

$$m(F(U)) = f(m(U)) = f(u), \qquad m(F(V)) = f(m(V)) = f(v), \tag{5.15}$$

and

$$\frac{1}{2}w(F(U)) = |f'(u)|\frac{1}{2}w(U) + \frac{1}{2!}|f''(u)|\left(\frac{1}{2}w(U)\right)^2 + \cdots + \frac{1}{n!}|F^{(n)}(U)|\left(\frac{1}{2}w(U)\right)^n, \tag{5.16}$$

with an analogous expression for $\frac{1}{2}w(F(V))$. In order to prove that $F(U) \subset F(V)$, it will be shown that $\frac{1}{2}w(F(U)) + |m(F(U)) - m(F(V))| \leq \frac{1}{2}w(F(V))$, from which the desired result follows by the Caprani–Madsen Lemma 4.1.

First, since $F^{(n)}$ is assumed to be monotone,

$$|F^{(n)}(U)|\left(\tfrac{1}{2}w(U)\right)^n \leq |F^{(n)}(V)|\left(\tfrac{1}{2}w(U)\right)^n. \tag{5.17}$$

Furthermore, by Theorem 5.1,

$$\begin{aligned} |f^{(k)}(u)| &\leq |f^{(k)}(v)| + \sum_{j=k+1}^{n-1} \frac{1}{(j-k)!}|f^{(j)}(v)| \cdot |u-v|^{j-k} \\ &\quad + \frac{1}{(n-k)!}|F^{(n)}(V)||u-v|^{n-k} \\ &= \sum_{j=k}^{n-1} \frac{1}{(j-k)!}|f^{(j)}(v)| \cdot |u-v|^{j-k} + \frac{1}{(n-k)!}|F^{(n)}(V)| \cdot |u-v|^{n-k}, \end{aligned} \tag{5.18}$$

$k = 1, 2, \cdots, n-1$, using the result of multiplication of (5.1) by $[-1, 1]$. It follows from (5.16), (5.17) and (5.18) that

$$\begin{aligned} \frac{1}{2}w(F(U)) \leq \sum_{k=1}^{n-1} \frac{1}{k!}\left\{\sum_{j=k}^{n-1} \frac{1}{(j-k)!}|f^{(j)}(v)||u-v|^{j-k}\right\}\left(\frac{1}{2}w(U)\right)^k \\ + \sum_{k=1}^{n} \frac{1}{k!(n-k)!}|F^{(n)}(V)||u-v|^{n-k}\left(\frac{1}{2}w(U)\right)^k. \end{aligned} \tag{5.19}$$

Interchange of order of the double summation in (5.19) results in

$$\begin{aligned} \sum_{k=1}^{n-1}\sum_{j=k}^{n-1} \frac{1}{k!(j-k)!}|f^{(j)}(v)| \cdot |u-v|^{j-k}\left(\frac{1}{2}w(U)\right)^k \\ = \sum_{j=1}^{n-1}\sum_{k=1}^{j} \frac{1}{k!(j-k)!}|f^{(j)}(v)| \cdot |u-v|^{j-k}\left(\frac{1}{2}w(U)\right)^k. \end{aligned} \tag{5.20}$$

Theorem 5.1 also gives

$$|f(u) - f(v)| \leq \sum_{j=1}^{n-1} \frac{1}{j!}|f^{(j)}(v)| \cdot |u-v|^j + \frac{1}{n!}|F^{(n)}(V)| \cdot |u-v|^n. \tag{5.21}$$

Addition of (5.19) and (5.21) results in the inequality

$$
\begin{aligned}
&\frac{1}{2}w(F(U))+|f(u)-f(v)| \\
&\le \sum_{j=1}^{n-1}\frac{1}{j!}|f^{(j)}(v)|\sum_{k=0}^{j}\frac{j!}{k!(j-k)!}|u-v|^{j-k}\left(\frac{1}{2}w(U)\right)^{k} \\
&\qquad +\frac{1}{n!}|F^{(n)}(V)|\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}|u-v|^{n-k}\left(\frac{1}{2}w(U)\right)^{k} \\
&=\sum_{j=1}^{n-1}\frac{1}{j!}|f^{(j)}(v)|\left\{\frac{1}{2}w(U)+|u-v|\right\}^{j}+\frac{1}{n!}|F^{(n)}(V)|\left\{\frac{1}{2}w(U)+|u-v|\right\}^{n}.
\end{aligned}
\tag{5.22}
$$

Hence, by the Caprani–Madsen Lemma 4.1,

$$
\begin{aligned}
&\frac{1}{2}w(F(U))+|m(F(U))-m(F(V))| \\
&\le \sum_{j=1}^{n-1}\frac{1}{j!}|f^{(j)}(v)|\left(\frac{1}{2}w(V)\right)^{j}+\frac{1}{n!}|F^{(n)}(V)|\left(\frac{1}{2}w(V)\right)^{n} \\
&=\frac{1}{2}w(F(V)),
\end{aligned}
\tag{5.23}
$$

and thus $F(U)\subset F(V)$. Q.E.D.

Remark 5.1. Some of the combinatorial aspects of the proof of Theorem 5.2 can be avoided by the use of the identity

$$
\begin{aligned}
&\phi(u)+\phi'(u)(x-u)+\cdots+\frac{1}{(n-1)!}\phi^{(n-1)}(u)(x-u)^{n-1} \\
&\qquad =\phi(v)+\phi'(v)\{(x-u)+(u-v)\}+\cdots \\
&\qquad\qquad +\frac{1}{(n-1)!}\phi^{(n-1)}(v)\{(x-u)+(u-v)\}^{n-1},
\end{aligned}
\tag{5.24}
$$

in which ϕ is an (abstract) polynomial of degree $n-1$ [12].

6. Application to iteration operators. The interval versions of the mean value and Taylor's theorem given above, like their counterparts in real and functional analysis, have numerous applications. Theorem 5.1 shows, for example, that the *interval remainder term*

$$
R_n(X)=\frac{1}{n!}F^{(n)}(X)\cdot(X-x)^{n} \tag{6.1}
$$

contains the *truncation error* $f(y)-f_{n-1}(y)$ resulting from the use of the *Taylor polynomial*

$$
f_{n-1}(y)=f(x)+f'(x)(y-x)+\cdots+\frac{1}{(n-1)!}f^{(n-1)}(x)(y-x)^{n-1} \tag{6.2}
$$

of degree $n-1$ in place of $f(y)$ for arbitrary $y\in X$. In particular, the results obtained by Moore [6], [7] on the numerical solution of differential equations by interval methods follow from this expansion.

The application to be considered here is to the solution of the equation

$$f(x)=0 \tag{6.3}$$

for $x\in D\to P$, where $f: D\subset P\to Q$ is a differentiable operator. Given a nonsingular linear operator $Y: Q\to P$, equation (6.3) can be transformed into the *fixed point problem* $x=g(x)$ for the operator $g: D\subset P\to P$ defined by

$$g(x)=x-Yf(x). \tag{6.4}$$

Since simple iteration is often used to solve fixed point problems, g will be called an *iteration operator* for f. The choice $Y=f'(x)^{-1}$ corresponds to *Newton's method* for the solution of (6.3), $Y=f'(z)^{-1}$ for $z\neq x$ to a method of *Stirling type* [11], and so on. Treating Y as a constant operator, one has

$$g'(x)=I-Yf'(x),\quad g''(x)=-Yf''(x),\quad \cdots,\quad g^{(n)}(x)=-Yf^{(n)}(x), \tag{6.5}$$

where I denotes the identity operator in P, and thus, if f is differentiable at least n times, then

$$\begin{aligned} g(x)\in{}& z-Yf(z)+\{I-Yf'(z)\}(x-z)-\cdots \\ & -\frac{1}{(n-1)!}Yf^{(n-1)}(z)(x-z)^{n-1}-\frac{1}{n!}YF^{(n)}(X)\cdot(X-z)^n \end{aligned} \tag{6.6}$$

for $x, z\in X$, where $F^{(n)}$ is an interval inclusion of $f^{(n)}$ on X.

Now, given an arbitrary sequence $Y_0, Y_1, \cdots$ of nonsingular linear operators, a sequence of intervals $X_0, X_1, \cdots$, and points $z_k\in X_k$, $k=0,1,2,\cdots$, one can construct the corresponding sequence $G_0, G_1, \cdots$ of *interval iteration operators* for f defined by

$$\begin{aligned} G_k(X_k)={}& z_k-Y_kf(z_k)+\{I-Y_kf'(z_k)\}\cdot(X_k-z_k)-\cdots \\ & -\frac{1}{(n-1)!}Y_kf^{(n-1)}(z_k)\cdot(X_k-z_k)^{n-1}-\frac{1}{n!}Y_kF^{(n)}(X_k)\cdot(X_k-z_k)^n, \end{aligned} \tag{6.7}$$

$k=0,1,2,\cdots$. The following theorem is a direct consequence of (6.6).

THEOREM 6.1. *If $x=x^*\in X_0$ is a solution of* (6.3), *then for*

$$X_{k+1}=X_k\cap G_k(X_k),\qquad k=0,1,2,\cdots, \tag{6.8}$$

one has

$$x^*\in X=\bigcap_{k=0}^{\infty}X_k. \tag{6.9}$$

Proof. It follows from (6.6) that $x^*\in X_k\Rightarrow x^*\in G_k(X_k)$, since $x^*=g(x^*)$, which in turn implies $x^*\in X_{k+1}$. This gives (6.9). Q.E.D.

The process (6.9) is called *interval iteration* [14]. Since

$$X_0\supset X_1\supset X_2\supset\cdots, \tag{6.10}$$

this process gives improved lower and upper bounds for x^* as long as $X_{k+1}\neq X_k$. The contrapositive of the assertion in Theorem 6.1 is that if

$$X_{k+1}=X_k\cap G_k(X_k)=\varnothing \tag{6.11}$$

for some positive integer k, where $\varnothing$ denotes the empty set, then $x^*\notin X_0$, and there is consequently no fixed point of g or solution of (6.3) in the initial interval X_0 [14].

In the case $n=1$, one obtains the *Krawczyk operators* [5]

$$K_k(X_k)=z_k-Y_kf(z_k)+\{I-Y_kF'(X_k)\}\cdot(X_k-z_k) \tag{6.12}$$

from (6.7), with $z_k=m(X_k)$. Suppose that F'' is an interval inclusion of f'' which is *consistent* with F' in the sense that

$$L-F'(X)\subset F''(X)w(X), \qquad L\in F'(X). \tag{6.13}$$

Then, from (6.12), for $Y_k^{-1}\in F'(X_k)$,

$$K_k(X_k)\subset z_k-Y_kf(z_k)+\tfrac{1}{2}Y_kF''(X_k)w(X_k)^2\cdot[-1,1], \tag{6.14}$$

since $X_k-z_k=X_k-m(X_k)=\frac{1}{2}w(X_k)\cdot[-1,1]$. It follows that interval iteration with the Krawczyk operator converges quadratically as $w(X_k)\to 0$ to a degenerate interval, thus mimicking the behavior of its real counterparts.

For $n=2$, the Chebyshev-type iteration operator [12]

$$T_k(X_k)=z_k-Y_kf(z_k)+\{I-Y_kf'(z_k)\}\cdot(X_k-z_k)-\tfrac{1}{2}Y_kF''(X_k)\cdot(X_k-z_k)^2 \tag{6.15}$$

results, and so on. It follows that (6.7) can be used to construct interval iteration operators with arbitrarily high orders of convergence in width as $w(X_k)\to 0$.

7. Other derivations of the mean value and Taylor's theorem. In certain particular cases, Taylor's theorem as given above (which includes the mean value theorem for $n=1$), can be derived directly from classical results in real or functional analysis. For example, with $P=Q=R$, one has

$$\begin{aligned} f(b)=f(a)+f'(a)(b-a)+\cdots+\frac{1}{(n-1)!}f^{(n-1)}(a)(b-a)^{n-1} \\ +\frac{1}{n!}f^{(n)}(\xi)(b-a)^n, \qquad a<\xi<b, \end{aligned} \tag{7.1}$$

in which the remainder term is said to be in *Lagrange* form. For $X=[a,b]$, one has $f^{(n)}(\xi)\in F^{(n)}(X)$, $b-a\in X-a$, which gives (5.1) at once in this special case. Formula (7.1) also holds componentwise in R^v, which leads to a similar generalization, since $f_k^{(n)}(\xi_k)\in F_k^{(n)}(X)$, $k=1,2,\cdots,v$, even though (7.1) does not necessarily hold for some $\xi\in X\subset R^v$. This generalization to R^v has been used by Moore [6], [7], and Caprani and Madsen [2]. In the latter paper, a version of the mean value theorem was also derived for integral operators, but the results are not easy to interpret without the use of interval integration [3], [13].

A more straightforward method of generalization of Taylor's theorem can be based on the use of the *Cauchy* form of the remainder term,

$$R_n(f;a,b)=\int_0^1 f^{(n)}(a+\theta(b-a))\frac{(1-\theta)^{n-1}}{(n-1)!}\cdot(b-a)^n d\theta, \tag{7.2}$$

which holds in Banach spaces [4], [12]. In R, the use of interval integration gives

$$\begin{aligned} R_n(f;a,b)&\in\int_0^1 F^{(n)}(a+\theta(b-a))\frac{(1-\theta)^{n-1}}{(n-1)!}\cdot(b-a)^n d\theta \\ &\subset F^{(n)}(X)\cdot(X-a)^n\int_0^1\frac{(1-\theta)^{n-1}}{(n-1)!}d\theta, \end{aligned} \tag{7.3}$$

from which (5.1) is obtained by evaluation of the real integral. By a simple extension of the concept of the interval integral [3], [13] to abstract functions f which take on values in a Banach space D, $D \subset Q$, Q a real space, a corresponding generalization of formula (7.3) will be obtained.

In order to construct the interval integral of an abstract function $\phi = \phi(\theta)$ which takes on values in an R-space Q for $0 \le \theta \le 1$, one simply partitions $\Theta = [0,1]$ into subintervals $\Theta_i = [\theta_{i-1}, \theta_i]$, $i = 1, 2, \cdots, m$, by means of points $0 = \theta_0 \le \theta_1 \le \cdots \le \theta_m = 1$. The set of all such partitions into m subintervals will be denoted by Δ_m. The *abstract interval function* $\Phi : I\Theta \to IQ$ will be defined by

$$\Phi(\Theta_i) = \left[\inf_{\Theta_i} \phi(\theta), \sup_{\Theta_i} \phi(\theta) \right]. \tag{7.4}$$

DEFINITION 7.1. The *interval integral* of the abstract function ϕ over $[0,1]$ is defined to be

$$\int_0^1 \phi(\theta)\, d\theta = \bigcap_{m=1}^{\infty} \bigcap_{\Delta_m} \sum_{i=1}^{m} \Phi(\Theta_i) w(\Theta_i) \in IQ. \tag{7.5}$$

This follows exactly the construction of [3]; again, the interval integral defined by (7.5) is the intersection of a nested sequence of nonempty intervals, and hence is nonempty.

Now, suppose that $D \subset Q$ is a Banach space in which $X \cap D$ is a closed set for $X \in IQ$. The Riemann (R) integral of abstract functions ϕ taking on values in D is defined to be the limit of the *Riemann sums*

$$\Sigma_{m,\Delta}\phi = \sum_{i=1}^{m} \phi(\tau_i)(\theta_i - \theta_{i-1}), \qquad \tau_i \in \Theta_i, \tag{7.6}$$

as $m \to \infty$ and $\|\Delta\| = \max_{(i)} w(\Theta_i) \to 0$ [4], [12]. It follows that

$$(R)\int_0^1 \phi(\theta)\, d\theta \in \sum_{i=1}^{m} \Phi(\Theta_i) w(\Theta_i) \subset \Phi(\Theta), \tag{7.7}$$

since the intersection of D with the *interval Darboux sums* [3] appearing in (7.5) is closed in the topology of D. Therefore, from (7.5),

$$(R)\int_0^1 \phi(\theta)\, d\theta \in \int_0^1 \phi(\theta)\, d\theta, \tag{7.8}$$

if ϕ is Riemann (R) integrable over $[0,1]$ in the sense defined by Graves [4]. Thus, in the special case that f is a function taking on values in a Banach space D with $f^{(n)}(a + \theta(b-a))$ Riemann integrable over $[0,1]$, (7.3) follows immediately by interval integration and gives (5.1) for interval inclusions $F^{(n)}$ of $f^{(n)}$. This derivation is also less general than the one given in §5, which holds in R-spaces.

REFERENCES

[1] G. BIRKHOFF, *Lattice Theory*, AMS Colloquium Publications, vol. 25, rev. ed., American Mathematical Society, New York, 1948.

[2] O. CAPRANI AND K. MADSEN, *Mean value forms in interval analysis*, Computing 25, (1980), pp. 147–154.

[3] O. CAPRANI, K. MADSEN AND L. B. RALL, *Integration of interval functions*, this Journal, 12 (1981), pp. 321–341.

[4] L. M. GRAVES, *Riemann integration and Taylor's theorem in general analysis*, Trans. Amer. Math. Soc., 29 (1927), pp. 163–177.

[5] R. KRAWCZYK, *Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken*, Computing, 4 (1969), pp. 187–201.
[6] R. E. MOORE, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[7] _____, *Methods and Applications of Interval Analysis*, SIAM Studies in Applied Mathematics, 2, Society for Industrial and Applied Mathematics, Philadelphia, 1979.
[8] K. NICKEL, *Verbandstheoretische Grundlagen der Intervall-Mathematik*, in [9], pp. 251–262 (1975).
[9] _____, ed., *Interval Mathematics*, Lecture Notes in Computer Science 29, Springer-Verlag, Berlin-Heidelberg-New York, 1975.
[10] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[11] L. B. RALL, *Convergence of Stirling's method in Banach spaces*, Aequationes Math., 12 (1975), pp. 12–20.
[12] _____, *Computational Solution of Nonlinear Operator Equations*, Wiley, New York, 1969; reprinted by Krieger, Huntington, NY, 1979.
[13] _____, *Integration of interval functions* II. *The finite case*, this Journal, 13 (1982), pp. 690–697.
[14] _____, *A theory of interval iteration*, Proc. Amer. Math. Soc., 86 (1982), pp. 625–631.
[15] W. SIERPIŃSKI, *General Topology*, C. Cecilia Krieger, tr., Univ. of Toronto Press, Toronto, 1952.
[16] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0003 $01.25/0

ON FUNCTIONS REPRESENTABLE AS A SUPREMUM OF A FAMILY OF SMOOTH FUNCTIONS*

Y. YOMDIN†

Abstract. Functions representable as a supremum of a bounded in a C^2-norm family of twice differentiable functions are considered. It is shown that the second generalized derivatives of such functions are measures. In particular, a new proof of this fact for convex functions follows. An upper estimate of the singular part of the second generalized derivatives is obtained in terms of some diameters of the family in the space of C^2-smooth functions. Applications to the distance function are given.

1. Introduction. We consider here the functions f representable as

$$f(x)=\sup_{h\in Q} h(x), \qquad x\in D\subset R^n, \tag{1.1}$$

where Q is a bounded in a C^2-norm family of twice differentiable functions on a domain $D\subset R^n$. It is shown in [7] that these functions (called in [7] subconvex) are rather similar to convex ones. In particular, the function f is subconvex if and only if it can be represented as

$$f=g_1-g_2, \tag{1.2}$$

where g_1 is convex, and g_2 is convex and C^2-smooth.

(Note that the high smoothness of functions h, without assumption of a boundness of Q in a C^2-norm, does not guarantee good properties of $f=\sup_{h\in Q}h$: e.g., any lower semicontinuous function can be represented as a supremum of a family of C^∞-smooth functions.)

From (1.2) it follows that the differentiability properties of subconvex functions are actually the same as those of convex ones. However, the representation (1.1) contains the information, which becomes less explicit in a form (1.2). (1.1) gives also a possibility to apply various analytical methods for the study of subconvex functions.

In this note we estimate the "measure of a nonsmoothness" of f in terms of some "diameters" of the set Q in a space of C^2-smooth functions.

The approach is based on the following two remarks: consider the finite set $\{h_1,\cdots,h_k\}$ of twice differentiable functions, and let $f=\max(h_1,\cdots,h_k)$. Typically f is a piecewise smooth function with "edges", on which the gradient ∇f has a jump. Define $\sigma(f)$ as the integral of the absolute value of this jump along all the edges. $\sigma(f)$ can serve the "measure of a nonsmoothness" of f. Then:

1. It turns out that $\sigma(f)$ is bounded by some constant, depending only on the common bound of the first and the second derivatives of $h_1,\cdots,h_k$, and not depending on the number k of the functions h_i involved and on their mutual position (while, say, the common length of the edges of f can be infinite).

2. The generalized second differential of f coincides with the usual one at smooth points of f and its singular part is equal to the integral along the edges of δ-functions, weighted by the jump of ∇f. Hence $\sigma(f)$ is closely related to the singular part of the generalized second differential of f. In turns out that $\sigma(f)$ is exactly the total variation of the singular part of a generalized Laplacian Δf.

* Received by the editors April 28, 1981, and in revised form April 22, 1982.
† Department of Mathematics, Ben Gurion University of the Negev, Beersheva 84120, Israel.

For a general subconvex function f, the total variation of a singular part of Δf is considered in this note as the measure of nonsmoothness of f (and is denoted by $\sigma(f)$). The main result (Theorem 2.3) is the upper estimate of $\sigma(f)$ by the size of Q in some special seminorm in a space of C^2-smooth functions.

Clearly, all the considerations in the case of subconcave function f ($f=\inf_{h\in Q}h$) are exactly the same.

Some applications to the distance function (which is one of the most important examples of subconcave functions) are given.

In all these results no assumptions on the structure of the family Q are used. If we consider $f=\max_t h(x,t)$, where $h(x,t)$ is a family of smooth functions smoothly depending on the finite-dimensional parameter t, the methods of singularity theory can be applied in a study of nonsmooth points of f. Some results in this direction are discussed in §5.

Remark. It can be shown that $\sigma(f)$ reflects the structure of the set $S_{n-1}(f)$ of points where the supporting hyperplanes of f at $(x,f(x))$ have one "degree of freedom" (see [1]). The analytical invariants reflecting the structure of the sets $S_k(f)$, $0\le k<n-1$ (see [1]) can also be built and estimated for subconvex functions. These results will appear separately.

2. Statement of the theorem. In what follows, m denotes the Lebesgue measure on R^n, s denotes the $(n-1)$-Lebesgue measure on smooth hypersurfaces in R^n, $x_1,\cdots,x_n$ are the standard coordinates in R^n, and $(\cdot,\cdot)$ and $\|\cdot\|$ denote the usual scalar product and the norm in R^n.

Let D be a closed bounded domain in R^n with a C^1-smooth boundary ∂D. The function $h: D\to R$ is said to be twice continuously differentiable in D if it can be extended to a twice continuously differentiable function $\hat{h}$ defined in some open neighborhood of D. Denote by $C^2(D)$ the space of all such functions with the norm

$$\|h\|_{C^2(D)}=\max_{x\in D}|f(x)|+\max_{x\in D}\sum_{i=1}^{n}\left|\frac{\partial h}{\partial x_i}(x)\right|+\max_{x\in D}\sum_{i,j=1}^{n}\left|\frac{\partial^2 h}{\partial x_i\partial x_j}(x)\right|.$$

Let $Q\subset C^2(D)$ be a subset bounded in the norm $\|\cdot\|_{C^2(D)}$. Let

$$K(Q)=\sup_{h\in Q}\|h\|_{C^2(D)},\qquad K_1(Q)=\sup_{h\in Q}\left(\max_{x\in D}|\Delta h(x)|\right).$$

Now for $h\in C^2(D)$ define the seminorms d_1, d_2 by

$$d_1(h)=\max_{y\in\partial D}|(\nabla h(y),n(y))|,\qquad d_2(h)=\max_{x\in D}|\Delta h(x)|.$$

Here ∇h is the vector of a gradient $(\partial h/\partial x_1,\cdots,\partial h/\partial x_n)$, Δh is the Laplacian $\sum_{i=1}^{n}\partial^2 h/\partial x_i^2$, and for $y\in\partial D$ $n(y)$ denotes the unit outward normal vector to ∂D at y.

Now we define $\delta(Q)$ as follows:

$$\delta(Q)=\inf_{g\in C^2(D)}\left(\sup_{h\in Q}d_1(h-g)s(\partial D)+\sup_{h\in Q}d_2(h-g)m(D)\right). \tag{2.1}$$

Taking $g=0$ in (2.1) we obtain

$$\delta(Q)\le K(Q)(s(\partial D)+m(D)). \tag{2.2}$$

To formulate the theorem we also need some notions of generalized functions and measure theory (see e.g. [4], [6]). Let $L(D)$ be the linear topological space of all the

C^∞-smooth functions with a compact support in $\mathring{D}$. The generalized function (or distribution) is a continuous linear functional on $L(D)$.

Each Borel measure μ on D (not necessarily positive) defines a distribution $[\mu]$ by

$$[\mu]\varphi=\int_{\mathring{D}}\varphi\,d\mu,\qquad \varphi\in L(D).$$

A necessary and sufficient condition for the distribution to be defined by the measure is given by the Riesz theorem: a distribution l is equal to $[\mu]$ for (uniquely defined) Borel measure μ if and only if there exists a constant K, such that

$$|l(\varphi)|\le K\cdot\max_{x\in D}|\varphi(x)|. \tag{2.3}$$

The total variation $|\mu|$ is then equal to the infinum of K, satisfying (2.3).

For the Borel measure μ let $\mu=\mu^r+\mu^s$ be the Lebesgue decompositon of μ. Here the regular part μ^r is given by the integral $\mu^r(A)=\int_A g\,dm$, $A\subset D$, and the singular part μ^s is concentrated on the set of Lebesgue measure zero.

For a locally integrable function f on D, the generalized Laplacian Δf is the distribution defined by $\Delta f(\varphi)=\int_D f\cdot\Delta\varphi$, $\varphi\in L(D)$.

Now let $Q\subset C^2(D)$ be a bounded subset, $f_Q(x)=\sup_{h\in Q}h(x)$, $x\in D$. It is well known that f_Q is a continuous and Lipschitzian function on D.

THEOREM 2.1. *The generalized Laplacian Δf_Q is defined by the measure $\mu(f_Q)$. The total variation $|\mu(f_Q)|$ does not exceed $K_1(Q)\cdot m(D)+\delta(Q)$.*

Let $\mu(f_Q)=\mu^r(f_Q)+\mu^s(f_Q)$ be the Lebesgue decomposition of $\mu(f_Q)$.

DEFINITION 2.2. The total variation $|\mu^s(f_Q)|$ is called the *measure of nonsmoothness of f_Q* and is denoted by $\sigma(f_Q)$.

THEOREM 2.3.

$$\sigma(f_Q)\le\delta(Q).$$

Remark. In a manner completely similar to the proof of Theorems 2.1 and 2.3, it can be shown that any generalized second partial derivative of f_Q is defined by a measure, and the total variation of the singular part of this measure does not exceed $\delta(Q)$.

Note that the representability of the generalized second derivatives of a convex function by a measure is well known (see [2]) and follows by (1.2) for subconvex functions. We obtain, therefore, as a by-product, another proof of this result.

3. Proof of theorems. First we consider the case where f is a maximum of a finite number of smooth functions. Let $h_1,\cdots,h_k\in C^2(D)$. We assume that these functions are in "general position", i.e.,

1. For each $i,j\le k$ the set $\{h_i-h_j=0\}$ is a regular submanifold in D $(i\ne j)$.

2. All the intersections of these submanifolds are transversal, as are their intersections with the boundary ∂D.

Note that if the given functions $h_1,\cdots,h_k$ are not in general position then, for any $\varepsilon>0$, we can find functions $h_1',\cdots,h_k'\in C^2(D)$ in general position such that $\|h_i'-h_i\|_{C^2(D)}\le\varepsilon$. (See e.g. [5].)

Let $Q=\{h_1,\cdots,h_k\}$, $f=f_Q=\max(h_1,\cdots,h_k)$. For each $i=1,\cdots,k$ denote by D_i the set $\{x\in D/f(x)=h_i(x)\}$, and let $\Gamma_{ij}=D_i\cap D_j$, $\Gamma_l=\partial D\cap D_l$. Clearly, each Γ_{ij} is a subregion of the submanifold $\{h_i-h_j=0\}$, and each D_i is a closed domain with $\partial D_i=\bigcup_{j\ne i}\Gamma_{ij}\cup\Gamma_i$. (Note that some of D_i,Γ_{ij},Γ_l could be empty or disconnected.) For $y\in\partial D_i$ denote by $n_i(y)$ the (a.e. defined) unit outward normal vector to ∂D_i at y.

Now let φ be a twice differentiable function on D. We have

$$\int_D f\cdot\Delta\varphi\,dm=\sum_{i=1}^{k}\int_{D_i}h_i\Delta\varphi\,dm. \tag{3.1}$$

To each integral in the right-hand side of (3.1) we apply the Green formula:

$$\int_{D_i}h_i\Delta\varphi\,dm=\int_{D_i}\Delta h_i\cdot\varphi\,dm-\int_{\partial D_i}(\nabla h_i,n_i)\cdot\varphi\,ds+\int_{\partial D_i}h_i(\nabla\varphi,n_i)\,ds. \tag{3.2}$$

Substituting the expressions (3.2) in (3.1) we obtain:

$$\int_D f\cdot\Delta\varphi\,dm=\sum_{i=1}^{k}\int_{D_i}\Delta h_i\cdot\varphi\,dm-\sum_{i,j=1}^{k}\int_{\Gamma_{ij}}\left[(\nabla h_i,n_i)+(\nabla h_j,n_j)\right]\varphi\,ds \tag{3.3}$$

$$+\sum_{l=1}^{k}\int_{\Gamma_l}\left[h_l(\nabla\varphi,n_l)-\varphi(\nabla h_l,n_l)\right]ds.$$

(Note that the integrals containing $\nabla\varphi$ are cancelled on each Γ_{ij}.)

Now, if $\varphi\in L(D)$ has a compact support in $\dot{D}$, the integrals over Γ_l in (3.3) vanish. Hence (3.3) gives a representation of a generalized Laplacian Δf by the measure. We see that the regular part of this measure is given by Δh_i on each D_i, and the singular part coincides with the Lebesgue $n-1$ measure on hypersurfaces Γ_{ij}, weighted by $-[(\nabla h_i,n_i)+(\nabla h_j,n_j)]$.

The following lemma shows that this weight is equal to the absolute value of the jump of ∇f on Γ_{ij}. Hence the singular measure $\mu^s(f)$ is positive, and its total variation $\sigma(f)$ coincides with $\sigma(f)$, defined in the introduction.

LEMMA 3.1.

$$-\left[(\nabla h_i,n_i)+(\nabla h_j,n_j)\right]=\|\nabla h_i-\nabla h_j\|$$

at each point of Γ_{ij}.

Proof. Since $n_i=-n_j$ at each point of Γ_{ij}, we have $(\nabla h_i,n_i)+(\nabla h_j,n_j)=(\nabla(h_i-h_j),n_i)$. Since Γ_{ij} is defined by $h_i-h_j=0$, the vector $\nabla(h_i-h_j)$ is collinear to n_i. Now, by definition of D_i, $f=\max(h_1,\cdots,h_k)$ is equal to h_i on D_i and is equal to h_j on D_j. Hence $h_i-h_j>0$ on D_i and $h_i-h_j<0$ on D_j, i.e., the direction of $\nabla(h_i-h_j)$ is opposite to the direction of n_i. Therefore

$$(\nabla(h_i-h_j),n_i)=-\|\nabla(h_i-h_j)\|. \tag{3.4}$$

□

COROLLARY 3.2.

$$\sigma(f)=\sum_{i,j}\int_{\Gamma_{ij}}\|\nabla h_i-\nabla h_j\|\,ds=-\sum_{i,j}\int_{\Gamma_{ij}}\left[(\nabla h_i,n_i)+(\nabla h_j,n_j)\right]ds.$$

The following lemma proves the estimate of Theorem 2.3 in a special case $f=\max(h_1,\cdots,h_k)$:

LEMMA 3.3.

$$\sigma(f)=\sum\int_{\Gamma_{ij}}\|\nabla h_i-\nabla h_j\|\,ds\le\delta(\{h_1,\cdots,h_k\}).$$

Proof. By (3.3), applied to $\varphi \equiv 1$,

$$\begin{aligned}\sigma(f) &= -\sum_{i,j=1} \int_{\Gamma_{ij}} \left[(\nabla h_i, n_i) + (\nabla h_j, n_j)\right] ds \\ &= \sum_{l=1}^{k} \int_{\Gamma_l} (\nabla h_l, n_l)\, ds - \sum_{i=1}^{k} \int_{D_i} \Delta h_i \cdot dm.\end{aligned} \tag{3.5}$$

Let $g \in C^2(D)$. For the smooth function g we have an equality:

$$0 = \int_{\partial D} (\nabla g, n)\, ds - \int_D \Delta g\, dm = \sum_{l=1}^{k} \int_{\Gamma_l} (\nabla g, n_l)\, ds - \sum_{i=1}^{k} \int_{D_i} \Delta g\, dm. \tag{3.6}$$

(Note that Γ_l form a partition of ∂D, and for $y \in \Gamma_l$, $n_l(y) = n(y)$.) Subtracting (3.6) from (3.5) we obtain:

$$\sigma(f) = \sum_{l=1}^{k} \int_{\Gamma_l} (\nabla h_l - \nabla g, n_l)\, ds - \sum_{i=1}^{k} \int_{D_i} (\nabla h_i - \nabla g)\, dm.$$

Now, for each $y \in \partial D$, $x \in D$,

$$|(\nabla h_l - \nabla g, n_l)|(y) \le \sup_k d_1(h_k - g), \qquad |\Delta h_i - \Delta g|(x) \le \sup_k d_2(h_k - g);$$

hence $\sigma(f) \le \sup_k d_1(h_k - g) \cdot s(\partial D) + \sup_k d_2(h_k - g) \cdot m(D)$, and taking inf over all $g \in C^2(D)$ we obtain the required inequality. □

Now turn back to the general situation. Let $Q \subset C^2(D)$ be a bounded subset, $f_Q = \sup_{h \in Q} h$.

PROPOSITION 3.4. *Let $\varphi \in L(D)$. Then*

$$\left| \int_D f_Q \cdot \Delta\varphi\, dm \right| \le \max_{x \in D} |\varphi(x)| \cdot \left[m(\operatorname{supp}\varphi) K_1(Q) + \delta(Q) \right],$$

where $m(\operatorname{supp}\varphi)$ is the Lebesgue measure of the support of φ.

Proof. Take some $\varepsilon > 0$. We can chose a finite number of functions $h'_1, \cdots, h'_k \in Q$ and then replace them by the functions $h_1, \cdots, h_k$ in a general position in such a way that for $f = \max(h_1, \cdots, h_k)$, on the one hand $|\int_D f_Q \cdot \Delta\varphi\, dm| \le |\int_D f \cdot \Delta\varphi\, dm| + \varepsilon$, and on the other hand $K_1(\{h_1, \cdots, h_k\}) \le K_1(Q) + \varepsilon$, $\delta(\{h_1, \cdots, h_k\}) \le \delta(Q) + \varepsilon$. But for f we have by (3.3) and Lemmas 3.1, 3.2 and 3.3:

$$\begin{aligned}\left| \int_D f \cdot \Delta\varphi\, dm \right| &\le \max_{x \in D} |\varphi(x)| \cdot \left(\int_{\operatorname{supp}\varphi} \max |\Delta h_i|\, dm + \sum_{i,j=1}^{k} \int_{\Gamma_{ij}} \|\nabla h_i - \nabla h_j\|\, ds \right) \\ &\le \max_{x \in D} |\varphi(x)| \left(m(\operatorname{supp}\varphi) \cdot K_1(\{h_1, \cdots, h_k\}) + \delta(\{h_1, \cdots, h_k\}) \right).\end{aligned}$$

Hence

$$\begin{aligned}\left| \int_D f_Q \cdot \Delta\varphi\, dm \right| &\le \left| \int_D f \cdot \Delta\varphi\, dm \right| + \varepsilon \\ &\le \max_{x \in D} |\varphi(x)| \left[m(\operatorname{supp}\varphi) \cdot (K_1(Q) + \varepsilon) + \delta(Q) + \varepsilon \right] + \varepsilon.\end{aligned}$$

Since ε can be chosen arbitrarily small, the proposition follows. □

Proof of Theorem 2.1. Since for each $\varphi \in L(D)$, $m(\operatorname{supp}\varphi) \le m(D)$, Proposition 3.4 shows that for the distribution Δf_Q the inequality (2.3) is satisfied with $K = m(D) \cdot K_1(Q) + \delta(Q)$. Hence by the Riesz theorem $\Delta f_Q = [\mu(f_Q)]$, and the total variation $|\mu(f_Q)|$ does not exceed $m(D) \cdot K_1(Q) + \delta(Q)$. □

To prove Theorem 2.3 let us consider the set $\Sigma \subset D$, on which the singular measure $\mu^s(f_Q)$ is concentrated, $m(\Sigma)=0$. The total variation $|\mu^s(f_Q)|$ is equal to the total variation $|\mu(f_Q)|(\Sigma)$ of $\mu(f_Q)$ on Σ, which in turn is equal to $\inf|\mu(f_Q)|(V)$, V is open, $V \supset \Sigma$. (see e.g. [6]). Now, by the Riesz theorem and Proposition 3.4, $|\mu(f_Q)|(V) \leq m(V) \cdot K_1(Q)+\delta(Q)$ ($\operatorname{supp}\varphi \subset V$ in our case).

Since $m(\Sigma)=0$, the measure $m(V)$, $V \supset \Sigma$, can be arbitrarily small, and we obtain the required inequality $\sigma(f_Q)=|\mu^s(f_Q)| \leq \delta(Q)$. □

Let us consider an example showing that the coefficients in the estimate of Theorem 2.3 cannot be improved. Let D be a closed disk in R^2 of radius r, centered at the origin.

Take $h_1=x_1^2+x_2^2$, $h_2=c$, $0 \leq c<r^2$, $f=\max(h_1,h_2)$. We have $\nabla h_1=(2x_1,2x_2)$, $\Delta h_1=4$, $\nabla h_2=0$, $\Delta h_2=0$. In computation of $\delta(\{h_1,h_2\})$ we can take $g=h_1/2$, and hence $\delta(\{h_1,h_2\}) \leq r \cdot 2\pi r+2 \cdot \pi r^2=4\pi r^2$.

The "edge" of the function f is the circle of the radius $\sqrt{c}$, and the jump of ∇f on this edge is equal to $2\sqrt{c}$. Hence $\sigma(f)=2\pi\sqrt{c} \cdot 2\sqrt{c}=4\pi c$, and for c tending to r^2, $\sigma(f)$ tends to the right-hand side of the inequality.

4. Applications to the distance function. It is convenient to consider the square of the usual distance function.

Let F be a closed subset in R^n. For $x \in R^n$ let $\rho_F(x)=\min_{y \in F}\rho(y,x)$, where $\rho(y,x)=\|y-x\|^2$.

Let Q be a bounded closed set in R^n, and let D be a bounded domain with a C^1-smooth boundary. We identify the set Q with the subset $\{\rho(y,\cdot), y \in Q\} \subset C^2(D)$.

LEMMA 4.1. $\delta(Q) \leq 2r(Q) \cdot s(\partial D)$, *where* $r(Q)$ *is the radius of the circumscribed ball of* Q *in* R^n.

Proof. We have $\nabla_x \rho(y,x)=2(x-y)$, $\Delta_x \rho(y,x) \equiv 2n$. Let y_0 be the center of a circumscribed ball of Q in R^n. Take $g=\rho(y_0,\cdot)$ in a definition of $\delta(Q)$. Since the difference of Laplacians $\Delta h-\Delta g$ is identically zero, and the difference $\|\nabla h-\nabla g\|=2\|(x-y)-(x-y_0)\|=2\|y_0-y\| \leq 2r(Q)$, then $\delta(Q) \leq 2r(Q) \cdot s(\partial D)$. □

Now let $F \subset R^n$ be an arbitrary closed subset D as above. Define $Q_F(D)$ as the set of all $y \in F$, which are the closest points in F to at least one point $x \in D$. By Lemma 4.1 and Theorem 2.3 we have:

PROPOSITION 4.2. *For the restriction* ρ_F *on* D,

$$\sigma(\rho_F/D) \leq 2r(Q_F(D)) \cdot s(\partial D).$$

In particular, for F bounded we obtain:

COROLLARY 4.3.

$$\sigma(\rho_F/D) \leq 2r(F) \cdot s(\partial D).$$

To formulate the last result here, we give the definition of the central set $C(F)$. A closed ball B, $\mathring{B} \subset R^n \setminus F$, which is not a proper subset of another ball B_1, $\mathring{B}_1 \subset R^n \setminus F$, is called a *maximal ball*. The set consisting of the centers of all maximal balls is called the *central set* $C(F)$ of F. (see e.g. [8]).

Let $x \notin C(F)$. Denote by B_η the closed ball of the radius η, centered at x.

COROLLARY 4.4. $\sigma(\rho_F/B_\eta)=O(\eta^n)$, *i.e.*, $\sigma(\rho_F/B)/\eta^n$ *remains bounded as* $\eta \to 0$.

Proof. It can be shown easily that for $x \notin C(F)$, $r(Q_F(B_\eta))=O(\eta)$ as $n \to 0$. □

Remark. Similar considerations show that for the usual distance function $\bar{\rho}_F=(\rho_F)^{1/2}$, the following inequality holds:

$$\sigma(\bar{\rho}_F/D) \leq \frac{r(Q_F(D))}{R^2} m(D)+\frac{r(Q_F(D))}{R} s(D);$$

R is the distance between the sets F and D.

5. The case of a smooth dependence on the parameter. In this section smooth means C^l differentiable, $l \geq 6$. Let $f(x)=\max_{t\in T} h(x,t)$, where $x\in D\subset R^n$, T is a compact smooth k-dimensional manifold and the function $h(x,t)$ is assumed to be smooth on (x,t). In this case the structure of the function f can be studied by the methods of singularity theory. The typical restriction here is that in order to obtain sufficiently detailed information, we should assume some kind of genericity of the family considered.

In applications this assumption is justified by the following facts:

1) The set of generic families is dense in the space of all families $h(x,t)$. In other words, any smooth family can be deformed to the generic one by an arbitrarly small perturbation.

2) Generic families are stable. In other words, if $h(x,t)$ is a generic family, then any family $\bar{h}(x,t)$ sufficiently close to $h(x,t)$ is generic and is equivalent, in some sense, to $h(x,t)$.

Thus the structure of $f(x)=\max_{t\in T} h(x,t)$ for $h(x,t)$ a generic family can be considered as a typical and a stable one.

We state here the theorem describing all the possible normal forms of f (up to addition of a smooth function and a smooth coordinate change) in a neigborhood of the point x_0 of a nonsmoothness, for a generic family $h(x,t)$ and the number n of variables x equal to $1,2,3$. (It turns out that the dimension k of the space T of parameters is not important in this description.) For each of these normal forms we give also the first terms of the development of $\sigma(\eta)=\sigma(f/B_\eta)$ by powers of η.

Let

$$\nu_1(x_1)=\max(x_1,-x_1)=|x_1|, \qquad \nu_2(x_1,x_2)=\max(x_1,x_2,-x_1-x_2),$$
$$\nu_3(x_1,x_2,x_3)=\max(x_1,x_2,x_3,-x_1-x_2-x_3), \qquad \kappa(x_1,x_2)=\max_t\left(-t^4+x_1t^2+x_2t\right).$$

Below, g will denote the smooth function. The precise definition of the notion of genericity used below can be found in [8].

THEOREM 5.1. *Let $f(x)=\max_{t\in T} h(x,t)$, where $x\in R^n$, $n=1,2,3$, and $h(x,t)$ is a generic family. Then in a neighborhood of each point x_0 of a nonsmoothness of f the coordinate system $x_1,\cdots,x_n$ can be chosen such that f in a neighborhood of x_0 can be written in one of the following forms*:

$n=1$. C_0^2: $f(x_1)=\nu_1(x_1)+g(x_1)$, $\sigma(\eta)=c$.

$n=2$. C_0^2: $f(x_1,x_2)=\nu_1(x_1)+g(x_1,x_2)$, $\sigma(\eta)=c\eta+O(\eta^2)$.
C_0^3: $f(x_1,x_2)=\nu_2(x_1,x_2)+g(x_1,x_2)$, $\sigma(\eta)=c\eta+O(\eta^2)$.
C_1^1: $f(x_1,x_2)=\kappa(x_1,x_2)+g(x_1,x_2)$, $\sigma(\eta)=c\eta^{3/2}+O(\eta^2)$.

$n=3$. C_0^2: $f(x_1,x_2,x_3)=\nu_1(x_1)+g(x_1,x_2,x_3)$, $\sigma(\eta)=c\eta^2+O(\eta^3)$.
C_0^3: $f(x_1,x_2,x_3)=\nu_2(x_1,x_2)+g(x_1,x_2,x_3)$, $\sigma(\eta)=c\eta^2+O(\eta^3)$.
C_0^4: $f(x_1,x_2,x_3)=\nu_3(x_1,x_2,x_3)+g(x_1,x_2,x_3)$, $\sigma(\eta)=c\eta^2+O(\eta^3)$.
C_1^1: $f(x_1,x_2,x_3)=\kappa(x_1,x_2)+g(x_1,x_2,x_3)$, $\sigma(\eta)=c\eta^{5/2}+O(\eta^3)$.
C_1^2: $f(x_1,x_2,x_3)=\max(\kappa(x_1,x_2),x_3)+g(x_1,x_2,x_3)$, $\sigma(\eta)=c_1\eta^2+c_2\eta^{5/2}+O(\eta^3)$.

The normal forms given here can be found (in slightly different notations) in [3]. For the topological description of the set of nonsmoothness of f see [8], where the

results are obtained in the case of the distance function. The description in the general case $f(x)=\max_t h(x,t)$ is exactly the same. The precise computations of $\sigma(\eta)$ in the forms given above, and in more complicated cases, will appear separately.

REFERENCES

[1] R. D. ANDERSON AND V. L. KLEE, *Convex functions and upper semicontinuous collections*, Duke Math. J., 19 (1952), pp. 349–357.

[2] I. BAKELMAN, *Geometric Methods of Solution of Elliptic Equations*, Moscow, 1965.

[3] L. N. BRYZGALOVA, *Singularities of a maximum of a function, depending on parameters*, Functional Anal. Appl., 11 (1977), N1, pp. 49–50.

[4] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, vol. I, Academic Press, New York, 1964.

[5] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Graduate Texts in Mathematics, New York, 1973.

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

[7] A. SHAPIRO AND Y. YOMDIN, *On functions representable as a difference of two convex functions, and necessary conditions in a constrained optimization*, preprint, 1981.

[8] Y. YOMDIN, *On the local structure of a generic central set*, Compositio Math., 43 (1981), pp. 225–238.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0004 $01.25/0

HILBERT TRANSFORM OF A FUNCTION HAVING A BOUNDED INTEGRAL AND A BOUNDED DERIVATIVE*

B. F. LOGAN†

Abstract. It is shown that if $f(x)$, $-\infty<x<\infty$, satisfies $|\int_a^b f(x)\,dx|\le M$ for any interval (a,b), and f is differentiable almost everywhere with $|f'(x)|\le m$, then $\tilde{f}$, the Hilbert transform of f, satisfies

$$|\tilde{f}(x)|\le\left(\frac{4}{\pi}\log 2\right)\sqrt{Mm},$$

and this inequality is best possible.

Suppose $f(\cdot)$ is any function defined on the real line satisfying

$$\left|\int_a^b f(x)\,dx\right|\le M \quad \text{for any interval } (a,b)$$

and f is differentiable almost everywhere with

$$|f'(x)|\le m.$$

Then $f(x)$ is bounded and so is its Hilbert transform.

THEOREM. *Under the above assumptions,*

$$\lim_{A\to\infty}\lim_{\varepsilon\to 0}\left|\int_{-A}^{-\varepsilon}+\int_{\varepsilon}^{A}\frac{f(t)}{\pi t}\,dt\right| \equiv\left|\int_{-\infty}^{\infty}\frac{f(t)}{\pi t}\,dt\right|\le\left(\frac{4}{\pi}\log 2\right)\sqrt{Mm}.$$

Proof. Define

$$F(t)=\int_0^t f(x)\,dx, \qquad F(0)=0.$$

Then we have the following identity for arbitrary C and positive T:

$$\int_{-\infty}^{\infty}\frac{f(t)\,dt}{t}=\int_{-T}^{T}\left\{\log\left|\frac{T}{t}\right|-C(T-|t|)\right\}f'(t)\,dt + \int_{|t|>T}\frac{F(t)\,dt}{t^2}+\left(C-\frac{1}{T}\right)\{F(T)+F(-T)\}.$$

Thus,

$$\left|\int_{-\infty}^{\infty}\frac{f(t)\,dt}{t}\right|\le m\int_{-T}^{T}\left|\log\left|\frac{T}{t}\right|-C(T-|t|)\right|dt+\frac{2M}{T}+\left|C-\frac{1}{T}\right|2M.$$

Now if T and C are chosen such that

$$m\left(\frac{T}{2}\right)^2=M, \qquad C\frac{T}{2}=\log 2,$$

* Received by the editors March 8, 1982.
† Bell Laboratories, Murray Hill, New Jersey 07974.

equality will be obtained above for the function

$$f(t)=\begin{cases} mt, & |t|\leq\frac{T}{2}, \\ m(T-|t|)\operatorname{sgn} t, & \frac{T}{2}<|t|\leq T, \\ 0, & |t|>T. \end{cases}$$

This function satisfies the conditions and the resulting inequality as stated in the theorem is best possible.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0005 $01.25/0

EXTREMAL PROBLEMS FOR POSITIVE-DEFINITE BANDLIMITED FUNCTIONS. I. EVENTUALLY POSITIVE FUNCTIONS WITH ZERO INTEGRAL*

B. F. LOGAN†

Abstract. A function $f(t)$ is bandlimited to $[-\lambda,\lambda]$ if it is the restriction to the real line of an entire function of exponential type $\leq\lambda$. This class of functions includes all functions whose Fourier transforms vanish outside $[-\lambda,\lambda]$. A real-valued function is positive definite if its Fourier transform is nonnegative on the real line. Such a function is necessarily even. In this paper we consider even real-valued functions $f(t)$ with $\int_{-\infty}^{\infty} f(t)\,dt=0$ which are eventually nonnegative, i.e.,

(i) $$f(t)\geq 0 \quad \text{for } |t|\geq T,$$

and whose Fourier transforms are nonnegative in a neighborhood of the origin (in particular, positive-definite $f(t)$), and vanish outside $[-1,1]$. We show that (i) cannot hold for $T<3\pi$ unless $f(t)$ vanishes identically. On the other hand, (i) holds for $T=3\pi$ and the positive-definite function

(ii) $$f(t)=\frac{(\cos t/2)^2}{(1-t^2/\pi^2)(1-t^2/9\pi^2)}.$$

This result is established by applying the Poisson sum formula to $(a^2-t^2)f(t)$, $0<a\leq\pi$, after proving for more general eventually-positive $f(t)$ that for any positive ε

(iii) $$\int_{-\infty}^{\infty}(1-\cos xt)f(t)\,dt\leq 0, \qquad 0\leq x<\varepsilon,$$

implies

(iv) $$\int_{\infty}^{\infty} t^2|f(t)|\,dt<\infty.$$

1. Introduction. This paper is the first of a series of papers on various extremal problems concerning the behavior of positive-definite bandlimited functions (see [1], [2], this issue, pp. 253–257, 258–268).

A. M. Odlyzko posed the following problem, which arose in connection with the question whether Dedekind zeta functions $\zeta_k(s)$ of algebraic number fields k of high degree have complex zeros close to (but not on) the real axis. Suppose $f(t)\not\equiv 0$ is an even real-valued function whose Fourier cosine transform $F(x)$ satisfies

(1) $$F(x)=\int_{-\infty}^{\infty} f(t)\cos xt\,dt\begin{cases}\geq 0, & -1\leq x\leq 1,\\ =0, & |x|>1,\end{cases}$$

(2) $$F(0)=0,$$

and in addition

(3) $$f(t)\geq 0, \qquad |t|\geq T.$$

What is the smallest T (and f) for which (3) can hold subject to (1) and (2)?

We must have $f(0)>0$, so $f(t)$ must have at least two sign changes in $(0,T]$ in order to satisfy (2); viz.,

$$\int_{-\infty}^{\infty} f(t)\,dt=0.$$

We ask how soon this sign changing can be accomplished subject to the condition (1).

* Received by the editors April 15, 1972.
† Bell Laboratories, Murray Hill, New Jersey 07974.

The condition (3) implies that $F(x)$ cannot vanish over an interval $(-\varepsilon, \varepsilon)$, for then $f(t)$ would be a high-pass function which must change sign, roughly speaking, as often as $\cos \varepsilon t$. (See [1, Thm. 5.4.3, p. 32].) Since the integral of $f(t)$ is (conditionally) convergent and $f(t)$ is eventually nonnegative, we conclude that $f(t)$ is absolutely integrable. Hence, $F(x)$ is continuous and vanishes outside $(-1, 1)$. Under these conditions, the Poisson sum formula

$$\tau \sum_{-\infty}^{\infty} f(k\tau+\theta) = \sum_{-\infty}^{\infty} F\left(\frac{2\pi k}{\tau}\right) e^{i2\pi k\theta/\tau} \tag{4}$$

is valid, giving the quadrature formula

$$\tau \sum_{-\infty}^{\infty} f(k\tau+\theta) = \int_{-\infty}^{\infty} f(t)\,dt, \qquad 0<\tau\leq 2\pi. \tag{5}$$

In particular, under conditions (1), (2) and (3),

$$\sum_{-\infty}^{\infty} f(2k\pi) = 0, \tag{6}$$

and since $f(0)>0$, we conclude that (3) cannot hold for T as small as 2π. The example

$$f(t) = \frac{(\cos t/2)^2}{(1-t^2/\pi^2)(1-t^2/9\pi^2)}, \tag{7}$$

$$F(x) = \begin{cases} \dfrac{3\pi^2}{8}|\sin \pi x|^3, & -1<x<1, \\ 0, & |x|\geq 1, \end{cases}$$

shows that (1), (2), (3) can hold for $T=3\pi$. This, in fact, is the extremal function, even with condition (1) relaxed to requiring the Fourier transform to be nonnegative only in a neighborhood of the origin.

THEOREM. *Let $f(t)$ be an even real-valued function whose Fourier transform $F(x)$ satisfies*

(i) $$F(x) = \int_{-\infty}^{\infty} f(t)\cos xt\,dt = 0, \qquad |x|>1,$$

(ii) $$F(0) = 0,$$

(iii) $$F(x)\geq 0, \qquad |x|<\varepsilon,$$

for some positive ε. Then

(iv) $$f(t)\geq 0 \quad \text{for } |t|\geq T$$

cannot hold for $T<3\pi$, unless

(v) $$f(t)\equiv 0,$$

and can hold for $T=3\pi$ with $f(t)$ given in (7).

In other words, if $f(t)\not\equiv 0$ and (iv) holds for $T<3\pi$, with $F(x)$ vanishing for $x=0$ and $|x|>1$, then $F(x)$ is negative in a neighborhood of the origin.

2. Proof of the theorem. In order to establish the theorem we prove a simple but interesting lemma.

LEMMA. *Let $f(t)$ be an even real-valued function satisfying*

(i) $$f(t)\geq 0, \qquad |t|>T$$

for some finite positive T, and whose Fourier transform

$$F(x)=\int_{-\infty}^{\infty} f(t)\cos xt\,dt$$

satisfies

(ii) $$F(0)-F(x)\leq 0, \qquad |x|<\varepsilon$$

for some positive ε. Then

(iii) $$\int_{-\infty}^{\infty} t^2|f(t)|\,dt<\infty,$$

and consequently $F(x)$ has a continuous second derivative, with

(iv) $$F''(0)=-\int_{-\infty}^{\infty} t^2 f(t)\,dt\geq 0.$$

For example, if $F(x)$ behaves like $|x|^{\nu}$ near the origin, then $f(t)$ cannot be positive for all sufficiently large $|t|$ unless $\nu\geq 2$. It is rather remarkable that (i) and (ii) place such a strong smoothness condition on $F(x)$.

Proof of the lemma. We have

$$\frac{F(0)-F(x)}{x^2}=\int_{-T}^{T}\frac{1-\cos xt}{x^2}f(t)\,dt+\int_{|t|>T}\frac{1-\cos xt}{x^2}f(t)\,dt\leq 0$$

for $0\leq x<\varepsilon$. Now as we let $x\to 0$, $(1-\cos xt)/x^2\to t^2/2$ on compact sets. Thus the first integral tends to a finite limit as $x\to 0$. The second integral is nonnegative for all $0\leq x<\varepsilon$ and hence is bounded above for all $0\leq x<\varepsilon$, since the sum of the two integrals is nonpositive, i.e.,

$$\int_{|t|>T} t^2 f(t)\,dt=\int_{|t|>T} t^2|f(t)|\,dt<\infty.$$

Therefore,

$$\int_{-\infty}^{\infty} t^2|f(t)|\,dt<\infty.$$

Then

$$\lim_{x\to 0}\int_{-\infty}^{\infty}\frac{1-\cos xt}{x^2}f(t)\,dt=\int_{-\infty}^{\infty}\frac{t^2}{2}f(t)\,dt\leq 0,$$

and the lemma is proved. □

Proof of the theorem. According to the lemma and the hypotheses of the theorem, the function $t^2 f(t)$ belongs to L_1 and its Fourier transform vanishes outside $(-1,1)$. Also we have from the lemma

(8) $$\int_{-\infty}^{\infty} t^2 f(t)\,dt\leq 0.$$

Now we define the function

(9) $$g_a(t)=(a^2-t^2)f(t), \qquad a>0,$$

which belongs to L_1 and whose Fourier transform vanishes outside $(-1, 1)$. We have

$$\int_{-\infty}^{\infty} g_a(t)\,dt \geq 0 \quad \text{for } a>0. \tag{10}$$

Now we apply the quadrature formula (5) to g_a with $\tau=2a$, $\theta=a$, and $a\leq\pi$, to obtain

$$2a \sum_{-\infty}^{\infty} g_a\{(2k+1)a\} = \int_{-\infty}^{\infty} g_a(t)\,dt \geq 0 \qquad (0<a\leq\pi). \tag{11}$$

Then using (9) and the fact that f is even we have

$$\sum_{1}^{\infty} k(k+1) f\{(2k+1)a\} \leq 0 \qquad (0<a\leq\pi). \tag{12}$$

Now if we suppose

$$f(t)\geq 0, \qquad |t|\geq 3(\pi-\varepsilon), \tag{13}$$

for some positive ε, we conclude from (12) and (13) that

$$f\{(2k+1)a\}=0 \quad \text{for } k=1,2,3,\cdots,(\pi-\varepsilon)\leq a\leq\pi; \tag{14}$$

i.e, that the entire function f vanishes over intervals. Thus the assumption (13) implies

$$f(t)\equiv 0. \tag{15}$$

This, with the example (7), completes the proof of the theorem. □

Acknowledgment. The author would like to thank J. C. Lagarias for assistance in the preparation of this manuscript.

REFERENCES

[1] B. F. LOGAN, *Properties of high-pass functions*, Doctoral thesis, Dept. Electrical Engineering, Columbia University, New York, 1965.

[2] _____, *Extremal problems for positive-definite bandlimited functions*. II. *Eventually negative functions*, this Journal, this issue, pp. 253–257.

[3] _____, *Extremal problems for positive-definite bandlimited functions*. III. *The maximum number of zeros in an interval* $[0, T]$, this Journal, this issue, pp. 258–268.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0006 $01.25/0

EXTREMAL PROBLEMS FOR POSITIVE-DEFINITE BANDLIMITED FUNCTIONS. II. EVENTUALLY NEGATIVE FUNCTIONS*

B. F. LOGAN†

Abstract. A function is bandlimited to $[-\lambda,\lambda]$ if it is the restriction to the real line of an entire function of exponential type $\leq\lambda$. This class of functions includes all functions whose Fourier transforms vanish outside $[-\lambda,\lambda]$. A real-valued function is positive definite if its Fourier transform is nonnegative on the real line. Such a function is necessarily even. In this paper we consider even real-valued functions $f(t)$ bandlimited to $[-1,1]$. These are functions of the form

$$f(t)=\int_0^1 \cos xt\, dF(x),$$

where $dF(x)$ is a bounded Stieltjes measure. We suppose that $f(0)=\int_0^1 dF(x)=1$. We show that if $dF(x)\geq 0$ for $0\leq x\leq\varepsilon$ for some $\varepsilon>0$, then $f(t)$ can satisfy

$$f(t)\leq 0 \quad \text{for } |t|\geq T$$

if and only if $T\geq\pi$; and for $T=\pi$ if and only if $f(t)$ is the positive-definite function

$$f(t)=\frac{(\cos t/2)^2}{1-t^2/\pi^2}=\frac{\pi}{2}\int_0^1 (\sin\pi x)\cos xt\, dx.$$

1. Introduction. This paper is the second of a series of papers on various extremal problems concerning the behavior of positive-definite bandlimited functions (see [1], [2], this issue, pp. 249–252, 258–268).

Suppose $f(t)$ is a real-valued positive-definite function bandlimited to $[-1,1]$; i.e., a function of the form

$$\begin{aligned} f(t)&=\int_0^1 \cos xt\, dF(x), \qquad dF(x)\geq 0,\\ f(0)&=\int_0^1 dF(x)=1. \end{aligned} \tag{1}$$

In [1] we showed that if a function $f(t)$ of the form (1) satisfies

$$\int_{-\infty}^{\infty} f(t)\, dt=0,$$

with the integral conditional convergent, then it would satisfy

$$f(t)\geq 0 \quad \text{for } |t|\geq T$$

if and only if $T\geq 3\pi$.

Here we ask the related question: What is the smallest T for which a function of the form (1) can satisfy

$$f(t)\leq 0 \quad \text{for } |t|\geq T? \tag{2}$$

To answer the question we need only require that $F(x)$ in (1) be nondecreasing in a neighborhood of the origin.

* Received by the editors April 15, 1982.
† Bell Laboratories, Murray Hill, New Jersey 07974.

THEOREM. *Let $f(t)$ be a function of the form*

$$f(t)=\int_0^1 \cos xt\, dF(x), \qquad f(0)=1,$$

where

$$dF(x)\geq 0 \quad \text{for } 0\leq x\leq \varepsilon \text{ and } 0<\varepsilon\leq 1.$$

Then $f(t)$ can satisfy

$$f(t)\leq 0 \quad \text{for } |t|\geq T$$

if and only if $T\geq\pi$; and for $T=\pi$ if and only if $f(t)$ is the positive definite function

$$f(t)=\frac{(\cos t/2)^2}{1-t^2/\pi^2}=\frac{\pi}{2}\int_0^1 (\sin \pi x)\cos xt\, dx.$$

Another interpretation of this result is: If a function of the form (1) has a first zero t_1 in the open interval $(0,\pi)$, then it must have a second sign change; otherwise we would have

$$f(t)<0 \quad \text{for } |t|>t_1,$$

in violation of the theorem. Note that $f(t)$ must be decreasing in $(0,\pi)$ (if $f(t)\not\equiv 1$, which the assumption $f(t_1)=0$ rules out); i.e.,

$$f'(t)=-\int_0^1 x\sin xt\, dF(x)<0 \quad \text{for } 0<t<\pi. \tag{3}$$

Hence t_1 in $(0,\pi)$ must be a simple zero; i.e., $f(t)$ changes sign at t_1. Note further that we must have $t_1\geq\frac{\pi}{2}$, since

$$f(t)=\int_0^1 \cos xt\, dF(x)\geq \cos t, \qquad \leq t\leq\pi, \tag{4}$$

i.e., $\cos xt$ takes its minimum value over $[0,1]$ at $x=1$ in case $0\leq t\leq\pi$. It would be interesting to determine how remote t_2, the next point of sign change, can be, given t_1 in $(\frac{\pi}{2},\pi)$. We leave this investigation for a future paper, giving the example

$$f_n(t)=\left(\sin\frac{\pi}{4n+2}\right)^2\frac{\{\cos((2n+1)t/(4n+1))\}^2\cos(t/(4n+1))}{\sin(\pi/(4n+2)-t/(4n+1))\sin(\pi/(4n+1)+t/(4n+1))} \tag{5}$$

where $t_1=\pi-\pi/(4n+2)$, $t_2=(4n+1)\pi/2$, $(n=1,2,\cdots)$. Note that

$$\lim_{n\to\infty} f_n(t)=\frac{\cos^2 t/2}{1-t^2/\pi^2}. \tag{6}$$

2. Proof of the theorem. We now turn to the proof of the theorem. We first observe that $F(x)$ cannot be constant over an interval including the origin, for then $f(t)$ would be a high-pass function which must have an infinite number of sign changes on a half-line [3].

LEMMA. *Suppose $f(t)$ is a real-valued function of the form*

$$f(t)=\int_0^\infty \cos xt\, dF(x), \qquad \int_0^\infty |dF(x)|<\infty$$

where F satisfies, for some positive ε,

$$F(x)-F(y)\geq 0, \qquad 0\leq y<x\leq\varepsilon.$$

Suppose also that f satisfies, for some positive T,

$$f(t)\leq 0, \qquad |t|\geq T.$$

Then f is absolutely integrable; hence

$$f(t)=\int_0^\infty \cos xt\,\phi(x)\,dx,$$

where $\phi(x)$ is continuous and absolutely integrable, and

$$\phi(x)\geq 0, \qquad 0\leq x\leq \varepsilon.$$

Hence

$$\int_{-\infty}^{\infty} f(t)\,dt\geq 0.$$

Proof of the lemma. We have

$$|f(t)|\leq\int_{-\infty}^{\infty}|\cos xt||dF(x)|\leq\int_{\infty}^{\infty}|dF(x)|=M<\infty. \tag{7}$$

Then if h is a function of L_1 we may invert the order of integration in

$$\int_{-\infty}^{\infty} f(t)h(t)\,dt=\int_{-\infty}^{\infty} h(t)\,dt\int_0^\infty \cos xt\,dF(x)$$

to obtain

$$\int_{-\infty}^{\infty} f(t)h(t)\,dt=\int_0^\infty \hat{h}(x)\,dF(x) \tag{8}$$

where $\hat{h}(x)=\int_{-\infty}^{\infty} h(t)\cos xt\,dt$. In particular,

$$\int_{-\infty}^{\infty} f(t)g(at)\,dt=\frac{\pi}{a}\int_0^{2a}\left(1-\frac{x}{2a}\right)dF(x), \qquad a>0, \tag{9}$$

where $g(t)=\{\sin t/t\}^2$. Since $F(x)$ is nondecreasing in $[0,\varepsilon]$ we have

$$\int_{-T}^{T} f(t)g(at)\,dt+\int_{|t|>T} f(t)g(at)\,dt\geq 0 \quad \text{for } 0<a<\varepsilon. \tag{10}$$

But $f(t)\leq 0$ for $t>T$; so

$$\int_{|t|>T} f(t)g(at)\,dt=-\int_{|t|>T}|f(t)|g(at)\,dt, \tag{11}$$

and hence from (10)

$$\int_{|t|>T}|f(t)|g(at)\,dt\leq\int_{-T}^{T} f(t)g(at)\,dt\leq\int_{-T}^{T}|f(t)|\,dt\leq 2MT \qquad (0<a<\varepsilon). \tag{12}$$

We have

$$g(at)\geq\frac{4}{\pi^2} \quad \text{for } |t|\leq\frac{\pi}{2a}. \tag{13}$$

Hence

$$\frac{8}{\pi^2}\int_T^{\pi/2a}|f(t)|\,dt\leq\int_{|t|>T}|f(t)|g(at)\,dt\leq\int_{-T}^{T}|f(t)|\,dt\leq 2MT \tag{14}$$

$$(0<a<\varepsilon),\ (a<\pi/2T).$$

Letting $a \to 0$ we obtain

$$\int_{-\infty}^{\infty} |f(t)|\, dt \leq \left(1+\frac{\pi^2}{4}\right) 2MT. \tag{15}$$

The cosine transform of $f(t)$ then is a continuous function. Thus

$$f(t) = \int_0^{\infty} \cos xt\, \phi(x)\, dx, \tag{16}$$

where $\phi(x)$ is continuous. Evidently

$$\phi(x) = \frac{dF(x)}{dx} \tag{17}$$

and

$$\int_0^{\infty} |\phi(x)|\, dx = \int_0^{\infty} |dF(x)| = M < \infty. \tag{18}$$

Also, since $F(x)$ is nondecreasing in $[0, \varepsilon]$, we have

$$\phi(x) \geq 0 \quad \text{for } 0 \leq x \leq \varepsilon. \tag{19}$$

Thus

$$\int_{-\infty}^{\infty} f(t)\, dt \geq 0 \tag{20}$$

and the lemma is proved. □

Under the hypotheses of the theorem, we have from the lemma

$$f(t) = \int_0^1 \phi(x) \cos xt\, dx, \tag{21}$$

where f belongs to L_1 and

$$\int_{-\infty}^{\infty} f(t)\, dt \geq 0. \tag{22}$$

Applying the Poisson sum formula to $f(t)$, we have

$$2\tau \sum_{-\infty}^{\infty} f\{(2k+1)\tau\} = \int_{-\infty}^{\infty} f(t)\, dt \geq 0 \qquad (0 < \tau \leq \pi). \tag{23}$$

Now suppose f satisfies

$$f(t) \leq 0 \quad \text{for } |t| \geq T, \quad \text{where } T < \pi. \tag{24}$$

We then conclude from (23) and (24) that

$$f\{(2k+1)\tau\} = 0 \quad \text{for } T \leq \tau \leq \pi \text{ and } k = 0, \pm 1, \pm 2, \cdots, \tag{25}$$

which, since $f(z)$ is an entire function, implies

$$f(t) \equiv 0, \tag{26}$$

which contradicts the assumption $f(0) = 1$. Thus (24) is false.

However, we can have (24) holding with $T = \pi$ as shown by the example

$$f(t) = \frac{\{\cos t/2\}^2}{1 - t^2/\pi^2}, \qquad f(t) = \frac{\pi}{2} \int_0^1 \sin \pi x \cos xt\, dx. \tag{27}$$

It remains to prove uniqueness. If f_1 is any function satisfying the hypotheses of the Theorem and

$$f_1(t)\leq 0, \qquad |t|\geq\pi, \tag{28}$$

we must have f_1 in L_1 (by the Lemma), and hence from (23)

$$f_1\{\pm(2k+1)\pi\}=0, \qquad k=0,1,2,\cdots, \tag{29}$$

and

$$f_1'\{\pm(2k+1)\pi\}=0, \qquad k=1,2,3,\cdots. \tag{30}$$

Also $f_1(z)$ is an entire function of exponential type 1 satisfying (cf. [4])

$$\int_{-\infty}^{\infty}|f_1(x+iy)|\,dx=o\{e^{|y|}\}, \qquad \lim_{x\to\pm\infty} f_1(x+iy)=0. \tag{31}$$

Hence

$$\lim_{y\to\pm\infty}\int_{-\infty}^{\infty}\frac{f_1(x+iy)\,dx}{f(x+iy)(x+iy)^2}=0, \tag{32}$$

where f is given by (27). It follows from (31) and (32) that

$$\lim_{n\to\infty}\int_{C_n}\frac{f_1(x)\,dz}{z(z-t)f(z)}=0 \tag{33}$$

where C_n is the contour formed by the lines

$$x=\pm 2n\pi, \qquad y=\pm 2n\pi.$$

Assuming in (33) that t is not a zero of f, and $t\neq 0$, we have, by the residue theorem, recalling (29) and (30),

$$\frac{f_1(0)}{-tf(0)}+\frac{f_1(t)}{tf(t)}=0. \tag{34}$$

Then since $f(0)=f_1(0)=1$ we have

$$f_1(t)=f(t), \tag{35}$$

which then holds for all t, since f_1 vanishes at the zeros of f. This completes the proof of the theorem.

Acknowledgments. The author would like to thank J. C. Lagarias for assistance in the preparation of this paper.

REFERENCES

[1] B. F. LOGAN, *Extremal problems for positive-definite bandlimited functions. I. Eventually positive functions with zero integral*, this Journal, this issue, pp. 249–252.

[2] _____, *Extremal properties for positive-definite bandlimited functions. III. The maximum number of zeros in an interval* [0, T], this Journal, this issue, pp. 258–268.

[3] _____, *Properties of high-pass functions*, Doctoral thesis, Dept. Electrical Engineering, Columbia University, New York, 1965.

[4] _____, *Limits in L_p of convolution transforms with kernels $aK(at)$, $a\to 0$*, this Journal, 10 (1979), pp. 733–740.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0007 $01.25/0

EXTREMAL PROBLEMS FOR POSITIVE-DEFINITE BANDLIMITED FUNCTIONS. III. THE MAXIMUM NUMBER OF ZEROS IN AN INTERVAL [0, T]*

B. F. LOGAN†

Abstract. A function is bandlimited to $[-\lambda,\lambda]$ if it is the restriction to the real line of an entire function of exponential type $\leq\lambda$. This class of functions includes all functions whose Fourier transforms vanish outside $[-\lambda,\lambda]$. A real-valued function is positive definite if its Fourier transform is nonnegative on the real line. Such a function is necessarily even. In this paper we consider even real-valued positive definite functions bandlimited to $[-1,1]$. These are functions of the form

$$f(t)=\int_0^1 \cos xt\, dF(x), \qquad dF(x)\geq 0,$$

with $dF(x)$ a bounded Stieltjes measure. We suppose that $f(0)=1$. Let $N(T)$ denote the number of zeros, counted according to multiplicity, in the closed interval $[0,T]$ of such a function $f(t)$. In this paper we show that

$$N(T)\leq\left[\!\left[\frac{2}{\pi}T\right]\!\right]$$

where $[\![x]\!]$ denotes the largest integer contained in x, with equality attaining for $T=\frac{n\pi}{2}$ if, and only if,

$$f(t)=\left(\cos\frac{t}{n}\right)^n;$$

and equality may attain in countless ways for $n\pi/2<T<(n+1)\pi/2$.

1. Introduction. A bounded bandlimited function $f(t)$, i.e., the restriction to the real line of an entire function of exponential type λ, may have an arbitrarily large number of zeros in a given interval: e.g.,

$$f(t)=P_{2n+1}(t)\,\frac{\sin\pi t}{t\prod_{k=1}^{n}(1-t^2/k^2)} \qquad (\lambda=\pi)$$

where P_{2n+1} is a polynomial of degree $(2n+1)$, say with $(2n+1)$ zeros in the given interval. However, if $f(t)$ has too many zeros (much more than $\cos\lambda t$) in the interval, then it can be shown that $f(t)$ must be relatively small over the interval. If f is even, type λ, $f(0)=1$, and satisfies

$$|f(t)|\leq f(0), \qquad -\infty<t<\infty, \tag{1}$$

then the number of zeros in [0, T] has some finite upper bound, depending on λT. It would be interesting to determine this upper bound, but here we restrict our attention to positive-definite bandlimited functions (type 1), i.e., functions of the form

$$f(t)=\int_{-1}^{1} e^{ixt}\,dF(x), \qquad dF(x)\geq 0. \tag{2}$$

We have

$$f(-t)=\overline{f(t)}. \tag{3}$$

* Received by the editors April 15, 1982, and in final form May 10, 1982.
† Bell Laboratories, Murray Hill, New Jersey 07974.

Hence f can have no more zeros in $[0,T]$ than does its real part; so for the problem at hand we may restrict our attention to functions of the form (see [1], [2], this issue, pp. 249–252, 253–257).

$$f(t)=\int_0^1 \cos xt\, dF(x), \qquad dF(x)\geq 0, \tag{4}$$

where for convenience we assume

$$f(0)=\int_0^1 dF(x)=1. \tag{4a}$$

We show the following result:

THEOREM. *Let $f(t)$ be any function of the form* (4), *and denote by $N(T)$ the number of zeros, counted according to multiplicity, of $f(t)$ in the closed interval* $[0,T]$. *Then*

$$N(T)\leq \left[\!\left[\frac{2}{\pi} T \right]\!\right]$$

where $[\![x]\!]$ denotes the largest integer contained in x. Equality is attained for $T=\frac{n\pi}{2}$, $n=1,2,3,\cdots$, if, and only if,

$$f(t)=\left\{\cos\frac{t}{n}\right\}^n,$$

and equality may attain in countless ways in case

$$\frac{n\pi}{2}<T<(n+1)\frac{\pi}{2}.$$

Note that for any fixed f we must have

$$\overline{\lim_{T\to\infty}} \frac{N(T)}{T}\leq\frac{1}{\pi}, \tag{5}$$

i.e., f cannot have, on the average, more zeros than $\cos t$, as is well known. Thus the upper bound for $N(T)$ cannot be achieved for all T by a fixed function f. We can think of the extremal functions for $[0,T]$ as gathering the "average allowable" number of zeros in $[0,2T]$ and placing them all at the center of the interval.

One approach to the problem is to assume

$$f(t_k)=0, \qquad k=1,2,\cdots,n, \tag{6}$$

$$0<t_1<t_2<t_3<\cdots<t_n=T. \tag{6a}$$

If n is sufficiently large, for fixed T one can find $\{a_k\}$ such that

$$\sum_1^n a_k \cos xt_k>0 \quad \text{for } 0\leq x\leq 1, \tag{7}$$

which contradicts the assumption (6) and gives

$$N(T)\leq n-1.$$

The difficulty here is in constructing sums satisfying (7) for *arbitrary* t_k in $[0,T]$ with the *minimal* (unknown) number n. We note from the example

$$f(t)=\prod_{k=1}^n \cos\lambda_k t, \qquad \sum_1^n \lambda_k\leq 1, \tag{8}$$

that $f(t_k)=0$, $k=1,2,\cdots,n$, is possible if we take

$$\lambda_k=\frac{\pi}{2t_k},$$

provided

$$\sum_1^n \lambda_k=\frac{\pi}{2}\sum_1^n \frac{1}{t_k}\leq 1. \tag{9}$$

Thus (7) cannot hold unless

$$\frac{\pi}{2}\sum_1^n \frac{1}{t_k}>1, \tag{10}$$

requiring

$$n>\frac{2t_1}{\pi}, \tag{11}$$

the worst case being $t_1\to T$. The theorem suggests that if (6a) is satisfied, then there should exist a sum satisfying (7), provided

$$n>\frac{2T}{\pi}. \tag{12}$$

This conclusion is probably correct, but here we use another approach to prove the theorem.

The crux of the proof here is to show that if $f(t)$ of the form (4) has n zeros in $[0,T]$ then there exists a function $f_n(t)$ of the same form which has an nth order zero at T. Then if f_n has an nth order zero at T it is relatively simple to show that $T\geq\frac{n\pi}{2}$.

2. Moving the zeros of certain positive-definite functions. Suppose f is a function of the form (4) with zeros t_k

$$0<t_1\leq t_2\leq t_3\leq\cdots\leq t_n=T.$$

Then it would be a simple matter to construct f_n having an nth order zero at T if removing the first zero t_1 of f resulted in a positive-definite function, i.e., if

$$g_{t_1}(t)=\frac{f(t)}{1-t^2/t_1^2}$$

were always positive definite. Then we could set

$$f_2(t)=\left(1-\frac{t^2}{T^2}\right)g_{t_1}(t)=\frac{t_1^2}{T^2}f(t)+\left(1-\frac{t_1^2}{T^2}\right)g_{t_1}(t)$$

which would also be positive definite. Then we could iterate the process, moving $\pm t_2$, the first zeros of f_2, to $\pm T$, etc., to obtain f_n with an nth order zero at $\pm T$. Unfortunately this is not always possible. We conjecture that it is if all the zeros of f are real.

Suppose f is positive definite and even. We consider

$$g_a(t)=\frac{f(t)-f(a)}{a^2-t^2} \tag{13}$$

and ask: under what conditions is g_a positive definite? We have

$$f(t)=\int_0^\infty \cos tx\,dF(x),\qquad dF(x)\geq 0, \tag{14}$$

and

$$g_a(t)=\int_0^\infty \frac{\cos tx-\cos ax}{a^2-t^2}\,dF(x). \tag{15}$$

In order to evaluate the Fourier transform of g_a, consider the integral

$$\begin{aligned}\phi(t)=\phi(t;\lambda,a)&=\frac{1}{a}\int_0^\lambda\{\sin a(\lambda-w)\}\cos wt\,dw \\ &=\frac{1}{2a}\int_0^\lambda \sin[a(\lambda-w)+wt]\,dw+\frac{1}{2a}\int_0^\lambda \sin[a(\lambda-w)-wt]\,dw \\ &=\frac{\cos t\lambda-\cos a\lambda}{a^2-t^2}.\end{aligned} \tag{16}$$

Thus

$$\int_0^\infty \frac{\cos t\lambda-\cos a\lambda}{a^2-t^2}\cos wt\,dt=\begin{cases}\dfrac{\pi}{2a}\sin a(\lambda-w), & 0<w\le\lambda,\\ 0, & w>\lambda.\end{cases} \tag{17}$$

We assume for the moment that $a>0$. Then

$$\int_0^\infty g_a(t)\cos wt\,dt=G_a(w) \tag{18}$$

where

$$G_a(w)=\frac{\pi}{2a}\int_w^\infty \sin a(x-w)\,dF(x), \qquad w\ge 0. \tag{18a}$$

So if $a>0$ we have

$$|G_a(w)|\le\frac{\pi}{2a}\int_0^\infty dF(x) \tag{19}$$

and

$$\lim_{a\to 0} G_a(w)=\frac{\pi}{2}\int_w^\infty (x-w)\,dF(x), \tag{20}$$

which is valid in case

$$\int_0^\infty x\,dF(x)<\infty. \tag{20a}$$

Now suppose f is any positive-definite function of the form

$$f(t)=\int_0^\lambda \cos xt\,dF(x), \qquad dF(x)\ge 0, \tag{21}$$

where $\int_0^\lambda dF(x)<\infty$ $(0<\lambda<\infty)$. Then (20a) holds and (20) gives

$$g_0(t)=\frac{f(0)-f(t)}{t^2}=\frac{\pi}{2}\int_0^\lambda G_0(x)\cos xt\,dx \quad \text{where } G_0(x)\ge 0. \tag{22}$$

Now we digress for a moment to give an example of a bandlimited positive-definite function f where t_1 is the first positive zero of f and

$$g_{t_1}(t)=\frac{f(t)}{t_1^2-t^2} \quad \text{is not positive definite.}$$

We take, using the result (22),

$$f(t)=A\,\frac{1-\sin\lambda t/\lambda t}{t^2}+\cos t, \qquad A>0, \tag{23}$$

where say, $\lambda=\frac{1}{2}$, and A is so large that the first positive zero t_1 of $f(t)$ satisfies

$$t_1>2\pi. \tag{24}$$

Now $f(t)$ has a spectral gap $(\frac{1}{2}, 1)$, and according to (18a) the cosine transform of $g_{t_1}(t)$ is

$$\begin{aligned} G_{t_1}(w)&=\frac{\pi}{2t_1}\int_w^\infty \sin t_1(x-w)\,dF(x)\\ &=\frac{\pi}{2t_1}\sin t_1(1-w) \quad \text{for } \tfrac{1}{2}\le w\le 1, \end{aligned} \tag{25}$$

which according to (24) will be negative somewhere in $(\frac{1}{2}, 1)$. Note that t_1 being large does not in itself prohibit g_{t_1} from being positive definite for other $f(t)$; e.g.,

$$f(t)=\frac{\sin\sqrt{t^2-a^2}}{\sqrt{t^2-a^2}}.$$

Note also from (18a) that if

$$f(t)=\int_0^1 \cos xt\,dF(x), \qquad dF(x)\ge 0,$$

and $f(t_1)=0$, $0<t_1<\pi$, then $g_{t_1}(t)=f(t)/(t_1^2-t^2)$ is also positive definite.

We return now to the question raised in (13). We have the following result:

LEMMA 1. *Suppose $f(t)$ is a function of the form*

i)

$$f(t)=\int_0^\infty \phi(x)\cos xt\,dx$$

where $\phi(x)$ is nonnegative and nonincreasing in $(0,\infty)$ and

ii)

$$\int_0^\infty \phi(x)\,dx<\infty.$$

Define for $a>0$

iii)

$$g_a(t)=\frac{f(t)-f(a)}{a^2-t^2}.$$

Then

iv)

$$g_a(t)=\frac{2}{\pi}\int_0^\infty G_a(x)\cos xt\,dx$$

where $G_a(x)\ge 0$ and $\int_0^\infty G_a(x)\,dx<\infty$. It follows then, if $f(a)=0$, that

$$\frac{b^2-t^2}{a^2-t^2}f(t)$$

is positive definite for $b\ge a$.

This lemma, of course, does not give the complete answer to the question raised. However, the result is sharp in the sense that, if $\phi(x)$ is not required to be decreasing, we can construct, as in the previous example, a function f for which the conclusion is false.

We note, since $g_a(t)$ is continuous, that $f(t)$ given by (i) must have a derivative for $t>0$. We have for g_a given by (iii)

$$g_a(a)=\frac{f'(a)}{-2a}, \qquad a>0, \tag{26}$$

and since $g_a(t)$ is a nonperiodic positive definite function, (assuming $g_a\not\equiv 0$)

$$|g_a(a)|<g_a(0)=\frac{f(0)-f(a)}{a^2}. \tag{27}$$

Thus for f of the form (i) ($f\not\equiv 0$) we have

$$|f'(a)|<2\,\frac{f(0)-f(a)}{a}, \qquad a>0. \tag{28}$$

Proof of Lemma 1. We observe that a nonincreasing nonnegative $\phi(x)$ $(0\le x<\infty)$ is a convex combination of the extreme functions

$$K_\lambda(x)=\begin{cases}\dfrac{1}{\lambda}, & 0\le x\le\lambda,\\ 0, & x>\lambda,\end{cases} \tag{29}$$

i.e.,

$$\begin{aligned}&\phi(x)=\int_0^\infty K_\lambda(x)\,d\mu(\lambda), \qquad d\mu\ge 0,\\ &\int_0^\infty \phi(x)\,dx=\int_0^\infty d\mu(\lambda),\end{aligned} \tag{30}$$

and hence

$$f(t)=\int_0^\infty \phi(x)\cos xt\,dx=\int_0^\infty \frac{\sin\lambda t}{\lambda t}\,d\mu(\lambda). \tag{31}$$

Thus it is sufficient to establish the lemma for

$$f(t)=f_\lambda(t)=\frac{\sin\lambda t}{\lambda t}. \tag{32}$$

In this case we have

$$g_a(t;\lambda)=\frac{\dfrac{\sin\lambda t}{\lambda t}-\dfrac{\sin\lambda a}{\lambda a}}{a^2-t^2}, \tag{33}$$

and according to (18), the cosine transform of g_a is

$$\begin{aligned}G_a(w;\lambda)&=\frac{\pi}{2a}\int_w^\infty \sin a(x-w)K_\lambda(x)\,dx\\ &=\begin{cases}\dfrac{\pi}{2a^2\lambda}\{1-\cos a(w-\lambda)\}, & 0\le w\le\lambda,\\ 0, & w>\lambda.\end{cases}\end{aligned} \tag{34}$$

Thus $G_a(w;\lambda)$ is nonnegative and

$$\frac{f(t)-f(a)}{a^2-t^2}=\int_0^\infty f_\lambda(t)\,d\mu(\lambda)=\frac{2}{\pi}\int_0^\infty G_a(w)\cos wt\,dt \tag{35}$$

where $G_a(w)=\int_0^\infty G_a(w;\lambda)\,d\mu(\lambda)\geq 0$ and

$$\frac{2}{\pi}\int_0^\infty G_a(w)\,dw=\frac{f(0)-f(a)}{a^2}<\infty \qquad (a>0). \tag{36}$$

Now if $f(a)=0$, we have

$$\frac{b^2-t^2}{a^2-t^2}f(t)=f(t)+\frac{b^2-a^2}{a^2-t^2}f(t)=f(t)+(b^2-a^2)g_a(t). \tag{37}$$

So if $b\geq a$, the function on the left in (37) is the sum of two positive-definite functions and is therefore itself positive definite. Thus if $f(t)$ is the cosine transform of a nonnegative nonincreasing function, any pair of zeros $(\pm a)$ of f may be moved outward to $(\pm b)$, the altered function being positive definite. This completes the proof of Lemma 1. □

3. Construction of f_n. We wish to prove the following:

LEMMA 2. *Suppose f is a function of the form*

$$f(t)=\int_0^1 \cos xt\,dF(x), \qquad dF(x)\geq 0,$$

$$\int_0^1 dF(x)>0,$$

having n zeros, counted according to multiplicity, in the closed interval $[0,T]$. Then there exists a function $f_n(t)$ of the same form having an nth order zero at T.

Proof. We may assume that the nth zero satisfies

$$t_n=T. \tag{38}$$

It simplifies the analysis to assume all the zeros are simple, and then let zeros coalesce to handle the general case. However, the construction of f_n depends on the order of the zero at T. So we assume that f has a ν^{th} order zero at T $(1\leq\nu<n)$ and indicate this by writing

$$f(t)=f_\nu(t). \tag{39}$$

To simplify the analysis we assume the remaining zeros are simple, i.e.,

$$0<t_1<t_2<\cdots<t_{n-\nu}<t_{n-\nu+1}=t_{n-\nu+2}=\cdots=t_n=T. \tag{40}$$

Then $f_\nu'(t)$ will have a zero of order $(\nu-1)$ at T and at least one zero t_k' between the separated zeros of f_ν, i.e.,

$$t_1<t_1'<t_2<t_2'<t_3<\cdots<t_{n-\nu}'<t_{n-\nu+1} \tag{41}$$

with $t_k'=T$ for $n-\nu+1\leq k\leq n-1$.

Now we wish to make a linear combination of $f_\nu(t)$ and

$$g_\nu(t)=\frac{T^2-t^2}{(t_1')^2-t^2}\left\{-\frac{f_\nu'(t)}{t}\right\}, \tag{42}$$

which will have a zero of order $(\nu+1)$ at T. First we observe that

$$-\frac{f_\nu'(t)}{t}=\int_0^1 \frac{\sin xt}{xt}\{x^2\,dF(x)\}, \tag{43}$$

which is a convex combination using the measure $x^2 dF(x)$ of functions $\sin xt/xt$ having nonnegative nonincreasing cosine transforms, as in (31). Hence by Lemma 1, $g_\nu(t)$ is positive definite. Also g_ν is a cosine integral of a function vanishing outside $[0, 1]$, according to (18a). Now we set

$$f_{\nu+1}(t)=f_\nu(x)+A_\nu g_\nu(t) \tag{44}$$

and determine A_ν so that $f_{\nu+1}$ has a zero of order $\nu+1$ at T. We have

$$f_\nu(T+S)=a_\nu S^\nu+a_{\nu+1}S^{\nu+1}+\cdots, \tag{45}$$

$$f_\nu'(T+S)=\nu a_\nu S^{\nu-1}+(\nu+1)a_{\nu+1}S^\nu+\cdots, \tag{46}$$

$$g_\nu(T+S)=\frac{-S(2T+S)}{(t_1')^2-(T+S)^2}-\frac{\left(-\nu a_\nu S^{\nu-1}+(\nu+1)a_{\nu+1}S^\nu+\cdots\right)}{T+S}. \tag{47}$$

Thus

$$f_{\nu+1}(T+S)=a_\nu S^\nu+A_\nu\frac{(2T)(-\nu a_\nu)S^\nu}{T^2-(t_1')^2}+O(S^{\nu+1}) \tag{48}$$

as $S\to 0$. So choosing

$$A_\nu=\frac{T^2-(t_1')^2}{2\nu T}>0 \tag{49}$$

gives $f_{\nu+1}$ a zero of order $(\nu+1)$ at T and $f_{\nu+1}$ is a positive-definite function of the same form as f_ν. We have

$$f_{\nu+1}(t_k)=A_\nu g_\nu(t_k)=\frac{T^2-t_k^2}{t_k^2-(t_1')^2}\frac{f_\nu'(t_k)}{t_k} \tag{50}$$

and

$$(-1)^k f_\nu'(t_k)>0 \quad \text{for } k=1,2,\cdots,n-\nu. \tag{51}$$

Hence, $f_{\nu+1}$ has a zero between each simple zero of f_ν, accounting for $n-\nu-1$ zeros, and then a zero of order $\nu+1$ at T; so $f_{\nu+1}$ is a function of the same form as f_ν with (at least) n zeros in $[0, T]$ and a zero of order $(\nu+1)$ at T. Thus the process can be iterated until f_n is obtained, having an nth order zero at T. This completes the construction of f_n and the proof of Lemma 2. □

4. The remoteness of an *n*th order zero. Now we prove the following:

LEMMA 3. *Suppose $f(t)$ is a function of the form*

$$f(t)=\int_0^1 \cos xt\, dF(x), \qquad dF(x)\geq 0,$$

with $\int_0^1 dF(x)=1$, having an nth order zero at $T(>0)$. Then

$$T\geq\frac{n\pi}{2},$$

with equality holding if, and only if,

$$f(t)=\left(\cos\frac{t}{n}\right)^n.$$

Proof. Suppose first that $n=2m$ is even. We then have

$$\int_{-1}^{1} x^{2k}\cos xT\,dF(x)=0, \qquad k=0,1,\cdots,m-1, \tag{52}$$

$$\int_{-1}^{1} x^{2k-1}\sin xT\,dF(x)=0, \qquad k=1,2,\cdots,m. \tag{53}$$

Now define

$$C_\nu(x)=(\cos x)\prod_{k=1}^{\nu}\left(1-\frac{4x^2}{\pi^2(2k-1)^2}\right), \tag{54}$$

$$S_\nu(x)=(x\sin x)\prod_{1}^{\nu}\left(1-\frac{x^2}{\pi^2k^2}\right). \tag{55}$$

We have

$$C_\nu(x)\geq 0 \quad \text{for } 0\leq x\leq(2\nu+1)\frac{\pi}{2}, \tag{56}$$

$$S_\nu(x)\geq 0 \quad \text{for } 0\leq x\leq(\nu+1)\pi. \tag{57}$$

and C_ν and S_ν have no common zeros.Then (52) and (53) imply

$$\int_0^1 C_{m-1}(xT)\,dF(x)=0, \tag{58}$$

$$\int_0^1 S_{m-1}(xT)\,dF(x)=0. \tag{59}$$

Now suppose $T\leq(2m-1)\frac{\pi}{2}$. Then both $C_{m-1}(xT)$ and $S_{m-1}(xT)$ are nonnegative for $0\leq x\leq 1$, and

$$S_{m-1}(xT)+C_{m-1}(xT)>0 \quad \text{for } 0\leq x\leq 1 \quad \text{and} \quad T\leq(2m-1)\frac{\pi}{2}. \tag{60}$$

So $T\leq(2m-1)\frac{\pi}{2}$ contradicts (58) and (59).

Next, suppose $(2m-1)\frac{\pi}{2}<T<m\pi$. Now $S_{m-1}(xT)$ is a nonnegative over $[0,m\pi/T]$ and positive in the open interval $((m-1)\pi/T,m\pi/T)$; whereas $C_{m-1}(xT)$ is nonnegative over $[0,(2m-1)\pi/2T]$ and negative in the open interval $((2m-1)\pi/2T,\ (2m+1)\pi/2T)$. It is clear, then, that for any T with $(2m-1)\pi/2<T<m\pi$ there is a sufficiently large A (depending on T) such that

$$AS_{m-1}(xT)+C_{m-1}(xT)>0 \quad \text{for } 0\leq x\leq 1, \tag{61}$$

which contradicts (58) and (59). This, with the previous contradiction, shows that $f(t)$ having a zero of order $2m$ at T implies

$$T\geq m\pi.$$

Now suppose $T=m\pi$. We have

$$\int_0^1 S_{m-1}(m\pi x)\,dF(x)=0 \tag{62}$$

and

$$S_{m-1}(m\pi x)\geq 0, \qquad 0\leq x\leq 1,$$

with equality for $x=\frac{k}{m}$, $k=0,1,\cdots,m$. Thus we must have

$$\frac{dF}{dx}=\sum_0^m a_k\delta\left(t-\frac{k}{m}\right), \qquad a_k\geq 0, \tag{63}$$

where $\Sigma_0^m a_k = 1$. Then according to (52) and (63)

$$\sum_0^m a_k P_{2m-2}\left(\frac{k}{m}\right)(\cos k\pi) = 0, \tag{64}$$

where $P_{2m-2}(x)$ is an even polynomial of degree $2m-2$ in x. Then a_j can be solved for, say, in terms of a_0 by choosing $P_{2m-2}(x)$ to vanish at $m-1$ of the $m+1$ points, $\frac{k}{m}$ $(k \neq j, 0)$. Then with $\Sigma_0^m a_k = 1$, the solution is unique. A solution, and hence the unique solution, is

$$\begin{aligned} a_0 &= \frac{1}{2^{2m}}\binom{2m}{m}, \\ a_k &= \frac{1}{2^{2m-1}}\binom{2m}{m-k}, \qquad k = 1, 2, \cdots, m. \end{aligned} \tag{65}$$

i.e., f can have a zero of order $2m$ at $T = m\pi$ if, and only if,

$$f(t) = \left(\cos\frac{t}{2m}\right)^{2m} \tag{66}$$

The proof of the lemma for odd order zeros is similar, with the roles of S_ν and C_ν interchanged. □

5. Proof of the theorem. Let $N(T)$ be the number of zeros, counted according to multiplicity, in $[0, T]$ of a function $f(t) = f(t; T)$ of the form

$$\begin{aligned} f(t) &= \int_0^1 \cos xt\, dF(x), \qquad dF(x) \geq 0, \\ f(0) &= \int_0^1 dF(x) = 1. \end{aligned} \tag{67}$$

We may suppose the Nth zero t_N occurs at T (replace $f(t)$ by $f(t \cdot t_N/T)$). Then, according to Lemma 2, there exists a function f_N of the form (67) having an Nth order zero at T. By Lemma 3 we must have

$$T \geq N\frac{\pi}{2}, \tag{68}$$

i.e.,

$$N(T) \leq \left[\!\left[\frac{2}{\pi}T\right]\!\right] \tag{69}$$

when $[\![x]\!]$ is the greatest integer contained in x. Again, from Lemma 3, equality may attain in (69) for $T = n\frac{\pi}{2}$, $N(T) = n$, if and only if,

$$f\left(t; \frac{n\pi}{2}\right) = \left(\cos\frac{t}{N}\right)^n. \tag{70}$$

For $n\pi/2 < T < (n+1)\pi/2$, equality in (69) may attain, for example, with

$$f(t; T) = \left(\cos\frac{\alpha t}{n}\right)^n g(t), \qquad \left(\frac{n\pi}{2} < T < (n+1)\frac{\pi}{2}\right) \tag{71}$$

where

$$\frac{n\pi}{2T} \leq \alpha < 1$$

and $g(t)$ is any positive-definite bandlimited function of type $(1-\alpha)$.

Another example is

$$f(t;T)=\prod_{k=1}^{n}\cos\lambda_k t, \qquad \frac{\pi}{2T}\leq\lambda_k\leq\frac{1}{n}, \tag{72}$$

where $n\pi/2<T<(n+1)\pi/2$.

This completes the proof of the theorem. □

Acknowledgment. The author would like to thank J. C. Lagarias for assistance in the preparation of this manuscript.

REFERENCES

[1] B. F. LOGAN, *Extremal problems for positive-definite bandlimited functions.* I. *Eventually positive functions with zero integral*, this Journal, this issue, pp. 249–252.

[2] _____, *Extremal problems for positive-definite bandlimited functions.* II. *Eventually negative functions*, this Journal, this issue, pp. 253–257.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0008 $01.25/0

AN INTEGRAL EQUATION CONNECTED WITH THE JACOBI POLYNOMIALS*

B. F. LOGAN†

Abstract. This paper considers the finite convolution equation

(i) $$\int_{-1}^{1} f(t)k(x-t)\,dt=g(x), \qquad -1<x<1,$$

where the kernel $k(t)$ is of the form

(ii) $$k(t)=\begin{cases} a|t|^{-\nu}, & x>0, \\ b|t|^{-\nu}, & x<0. \end{cases}$$

Here a and b are nonzero real numbers with $|a|\neq|b|$, and $\nu<1$ is not integral. The equation (i) can be reduced by successive differentiation or by one integration to the case where $k(x)$ is positive and $-1<\nu<1$, ($\nu\neq 0$). We show that in this case the solution to (i), if it exists, is unique. A closed form solution to (i) is obtained by considering the relation between the real and imaginary parts of the analytic function $G(z)$ defined by

(iii) $$G(z)=\frac{i}{\pi}e^{i\pi\alpha}\int_{-1}^{1} f(t)(z-t)^{-\nu}dt, \qquad \operatorname{Im} z\geq 0,$$

where $f(t)$ is assumed to be real valued. Certain moment relations between f and g in (i) are derived, leading to the result

(iv) $$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}P_n^{(\beta,\alpha)}(t)k(x-t)\,dt=c_nP_n^{(\alpha,\beta)}(x), \qquad -1<x<1,$$

where $P_n^{(\alpha,\beta)}$ is the Jacobi polynomial and the parameters α and β are simply related to a, b, and ν in (ii).

Another form of the solution to (i) is obtained by making use of some simple properties of the function defined for $-\infty<x<\infty$ by the integral in (iv) for the case $n=0$.

We discuss a number of other topics, including a conformal mapping characterization of certain functions $G(z)$ which are representable in the form (iii) with a real-valued function $f(t)$.

1. Introduction.

Finite convolution equations are integral equations of the form

$$f(x)=\int_{-a}^{a} k(x-y)g(y)\,dy$$

with $a<\infty$. They were considered by Carleman [1] in the case where $k(x)$ is either a logarithm or a power. Further work on this subject was done by Latta [3], and his method was generalized by Shinbrot [9]. (See Cochran [2, pp. 301–306] for a discussion of their work.) The kernel $k(t)=\frac{1}{t}$ gives the finite Hilbert transform. The integral equation

$$f(x)=\int_{-1}^{1}\frac{g(y)}{y-x}dy, \qquad -1<x<1,$$

is called the *airfoil equation*. It was analyzed by Tricomi [14]. (See also [15, pp. 173–185].) The kernel $k(t)=\ln|t|$ has a closed form inversion formula which Carleman [1] found (see [5, p. 203]).[1] Pearson [5] considered a more general kernel $k(t)=P(t)\ln|t|+Q(t)$ where $P(t)$ and $Q(t)$ are polynomials. There is no general solution method known for finite convolution equations, and most work has been done on equations with kernels $k(t)$ of a special form.

*Received by the editors April 16, 1982.

†Bell Laboratories, Murray Hill, New Jersey 07974.

[1] It appears as an exercise in Stakgold [11, p. 191].

In this paper we consider the integral equation

$$\int_{-1}^{1} f(t)k(x-t)\,dt=g(x), \qquad |x|<1, \tag{1.1}$$

where the kernel is of the special form

$$k(t)=a|t|^{-\nu}+b|t|^{-\nu}\operatorname{sgn} t, \qquad |t|>0, \tag{1.2}$$

with a and b real and $\nu<1$. We exclude the simple case[2] $|a|=|b|$. We say the kernel is of the *first type* when $|a|>|b|$ and of the *second type* when $|a|<|b|$. The corresponding integral equations will be referred to as the *first* and *second problems*, respectively. We take g to be defined on $(-\infty,\infty)$ by the extended definition of the kernel, while in the problem we are only given its projection on $(-1,1)$. We exclude the case when ν is an integer, as the equation can then be treated by elementary methods. For $\nu<0$, the equation can be differentiated a sufficient number of times to reduce it to the case $0<\nu<1$, each differentiation changing the type of the kernel. We will show that the solution to the equation, if it exists, is unique except when the kernel is of the second type with $0<\nu<1$. For this reason we solve the first problem for $-1<\nu<1$ ($\nu\neq 0$), so that if given the first problem with $-1<\nu<0$, one need not differentiate the equation to obtain the case when the solution is not unique.

Shinbrot [10], as an example of the application of his general Wiener–Hopf method, gave the solution to the first problem for the symmetric kernel, with $0<\nu<1$, in the form of a series of Gegenbauer polynomials. Here, employing entirely different methods, we give several forms of the solution of the problem for the general kernel of the first type ($-1<\nu<1$, $\nu\neq 0$), one being a series of Jacobi polynomials which in case of a symmetric kernel are proportional to the Gegenbauer polynomials. So Shinbrot's solution is actually valid for $-1<\nu<1$ ($\nu\neq 0$).

We assume at the onset that $f(t)$ in (1.1) is a function of L_1, although the extension to signed measures of the form $d\mu(t)$, where μ is a function of bounded variation, is no problem. Also with no loss in generality, we assume that f and g in (1.1) are real-valued functions. We note that owing to the special nature of the kernel, a bilinear transformation

$$t'=\frac{1+t}{1-t}, \qquad x'=\frac{1+x}{1-x} \tag{1.3}$$

carries (1.1) into an equation of the same form on $(0,\infty)$. However, it is more convenient to work with the interval $(-1,1)$, and the connection with the Jacobi polynomials is more direct.

In §2 we standardize the kernels in terms of two parameters α and β, anticipating the connection with the Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$, as well as making possible a convenient identification of $g(x)$ in (1.1) as the real part of a certain analytic function.

In §3 we solve the first problem for $-1<\nu<1$ ($\nu\neq 0$) by introducing the analytic function

$$G(z)=\frac{i}{\pi}e^{i\pi\alpha}\int_{-1}^{1} f(t)(z-t)^{-\nu}dt, \qquad \nu=\alpha+\beta+1, \quad \operatorname{Im} z\geq 0, \tag{1.4}$$

where we are given $g(x)=\operatorname{Re} G(x+i0)$ only in the interval $-1<x<1$. The trick then is to multiply $G(z)$ by an analytic function, viz., $(1-z)^{\alpha}(1+z)^{\beta}$, which is real in $(-1,1)$

[2] In this case the kernel $k(t)$ vanishes either for $t>0$ or $t<0$, and the solution is straightforward.

and such that the real part of the product

$$F(z)\equiv(1-z)^{\alpha}(1+z)^{\beta}G(z) \tag{1.5}$$

vanishes on the real axis outside [1, 1]. We then know the real part of $F(z)$ on the real axis and $F(z)=O(z^{-1})$ as $z\to\infty$ so that we can recover $F(z)$ from its real part on the real axis and hence recover $G(z)$ from the projection of its real part on $(-1,1)$. Then a certain linear combination $\gamma_{\nu}(x)$ of the real and imaginary parts of $G(x)$ is related to f in (1.4) by

$$\frac{1}{\pi}\sin\pi\nu\int_{-1}^{x}f(t)(x-t)^{-\nu}dt=\gamma_{\nu}(x), \tag{1.6}$$

which is readily solved for $f(t)$ knowing $\gamma_{\nu}(x)$ for $-1<x<1$.

In §4 we solve the second problem for $0<\nu<1$, the interpretation now being that we know the *imaginary* part of $G(z)$ defined by (1.4) only in the interval $(-1,1)$. We then multiply $F(z)$ defined in (1.5) by $\sqrt{1-z^2}$ to obtain

$$H(z)=\sqrt{1-z^2}\,F(z), \tag{1.7}$$

a function whose *imaginary* part vanishes outside $[-1,1]$ and hence we know its imaginary part on the real axis. The difficulty now is that

$$H(z)=O(1)\quad\text{as } z\to\infty. \tag{1.8}$$

In fact,

$$H(z)=\frac{1}{\pi}\int_{-1}^{1}f(t)\,dt+O(z^{-1})\quad\text{as } z\to\infty. \tag{1.9}$$

The upshot of this is that we can only determine $f(t)$ within a multiple of $(1-t)^{\beta-1/2}\cdot(1+t)^{\alpha-1/2}$. (In the second problem for $0<\nu<1$, the parameters α and β satisfy $\alpha>-\frac{1}{2}$, $\beta>-\frac{1}{2}$.)

In §5 we establish the connection of the Jacobi polynomials with the integral equation. First a result of §3 leads to the important formula (5.3), which implies certain moment relations between f and g and from which we deduce

$$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}P_n^{(\beta,\alpha)}(t)k_{\alpha,\beta}(x-t)\,dt=\frac{(\nu)_n}{n!}P_n^{(\alpha,\beta)}(x),\qquad -1<x<1, \tag{1.10}$$

where $k_{\alpha,\beta}(x)$ is the standard kernel with exponent $-\nu$ defined in §2 and $\nu=\alpha+\beta+1$. Note that if $\alpha=\beta$ then (1.10) shows $P_n^{(\alpha,\beta)}(x)$ is an eigenfunction of the finite convolution equation with symmetric kernel

$$k(t)=k_{\alpha,\alpha}(t)(1-t^2)^{\alpha},$$

and eigenvalue $(\nu)_n/n!$. We note that Rahman [7], [8] has given a 5-parameter family of kernels for which certain Jacobi polynomials are eigenfunctions. We have not examined the relation of this kernel to the kernels he constructs.

In §5 we also obtain certain necessary moment conditions between derivatives of g that must be satisfied, in order for

$$\int_{-1}^{1}f(t)(x-t)^{m}k_{\alpha,\beta}(x-t)\,dt=g(x),\qquad -1<x<1, \tag{1.11}$$

to have a solution, where m is a positive integer and $-1<\alpha<0$, $-1<\beta<0$. Equation (1.10) leads to a solution of the first problem in the form

$$f(t)=(1-t)^{\beta}(1+t)^{\alpha}\sum a_nP_n^{(\beta,\alpha)}(t).$$

In §6 an alternate integral solution to the first problem is developed, the key being the function

$$\sigma(x)=\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}k_{\alpha,\beta}(x-t)\,dt, \qquad -\infty<x<\infty, \tag{1.12}$$

which, according to (1.10), is constant in the interval $(-1,1)$, and as it turns out its derivative outside the interval is an elementary function. This, and the fact tat $k_{\alpha,\beta}(ax)=a^{-\nu}k_{\alpha,\beta}(x)$ for $a>0$, are used to derive a solution that is striking in that it does not resemble the form of the solution in §3.

In §7 a fairly simple connection is established between solutions of equations with conjugate kernels $k_{\alpha,\beta}(x)$ and $\tilde{k}_{\alpha,\beta}(x)$ with a common right-hand member $g(x)$.

In §8 certain simple linear operator pairs are established for f and its transform g under the convolution kernel $k_{\alpha,\beta}$.

In §9 we study the question of which functions $G(z)$ can be represented in the form (1.4). We show that if an analytic function $G(z)$ maps the upper half plane into a certain wedge-like region with rectifiable boundary, then it has the representation (1.4) where $f(t)$ is real valued. Some examples are given.

In §10 we discuss kernel expansions in Jacobi series such as

$$k_{\alpha,\beta}(x-t)=\sum_{0}^{\infty}a_nP_n^{(\alpha,\beta)}(x)P_n^{\alpha,\beta}(t). \tag{1.13}$$

The special case $\alpha=\beta$ of (1.13) is essentially an identity of Polya and Szegö [6, Hilfssatz 2, p. 27]:

$$|x-t|^{-\nu}=\frac{\Gamma\left(\frac{1+\nu}{2}\right)\Gamma\left(\frac{1-\nu}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\sum_{m=0}^{\infty}\left(1+\frac{2}{\nu}m\right)P_m^{(\nu/2)}(x)P_m^{(\nu/2)}(t).$$

R. Askey has pointed out to me that the methods of this paper may conceivably yield new results in potential theory; cf. [6]. In §10 we also give an expansion involving Gegenbauer polynomials (for the nonsymmetric kernel).

Finally, in §11 we apply the method of §6 to the solution of

$$\int_{-1}^{1}f(t)\log|x-t|\,dt=g(x), \qquad -1<x<1. \tag{1.14}$$

As we mentioned before, this equation was considered by Carleman [1]. The form of our solution bears no resemblance to that of the "analytic function" method and leads to an interesting identity for the finite Hilbert transform.

2. The standard kernels. We standardize the kernels by introducing the function analytic in the upper half plane:

$$K_{\alpha,\beta}(z)=\frac{1}{\pi}e^{i\pi(\alpha+1/2)}z^{-\nu}, \tag{2.1}$$

where $\nu=\alpha+\beta+1$, $0\le\arg z\le\pi$. On the real line we have

$$K_{\alpha,\beta}(x)=\begin{cases}\dfrac{1}{\pi}e^{i\pi(\alpha+1/2)}|x|^{-\nu}, & x>0,\\[2ex] \dfrac{1}{\pi}e^{-i\pi(\beta+1/2)}|x|^{-\nu}, & x<0.\end{cases} \tag{2.2}$$

We set

$$K_{\alpha,\beta}(x)=k_{\alpha,\beta}(x)+i\tilde{k}_{\alpha,\beta}(x), \tag{2.3}$$

where

$$k_{\alpha,\beta}(x)=\begin{cases}\dfrac{-\sin\pi\alpha}{\pi}|x|^{-\nu}, & x>0,\\ \dfrac{-\sin\pi\beta}{\pi}|x|^{-\nu}, & x<0 \quad (\nu=\alpha+\beta+1),\end{cases} \tag{2.4}$$

$$\tilde{k}_{\alpha,\beta}(x)=\begin{cases}\dfrac{\cos\pi\alpha}{\pi}|x|^{-\nu}, & x>0,\\ \dfrac{-\cos\pi\beta}{\pi}|x|^{-\nu}, & x<0 \quad (\nu=\alpha+\beta+1).\end{cases} \tag{2.5}$$

For the case $\alpha=\beta=\frac{1}{2}(\nu-1)$, we use a single subscript notation for the even and odd kernels; viz.,

$$k_{\alpha,\alpha}(x)\equiv k_{\nu}(x)=\frac{1}{\pi}\cos\frac{\pi\nu}{2}|x|^{-\nu}, \tag{2.6}$$

$$\tilde{k}_{\alpha,\alpha}(x)\equiv\tilde{k}_{\nu}(x)=\frac{1}{\pi}\sin\frac{\pi\nu}{2}|x|^{-\nu}\operatorname{sgn}x. \tag{2.7}$$

From the relation

$$e^{i\pi\lambda}K_{\alpha,\beta}(z)=K_{\alpha+\lambda,\beta-\lambda}(z) \tag{2.8}$$

we have

$$\cos\pi\lambda k_{\alpha,\beta}(x)-\sin\pi\lambda\tilde{k}_{\alpha,\beta}(x)=k_{\alpha+\lambda,\beta-\lambda}(x), \tag{2.9}$$

$$\sin\pi\lambda k_{\alpha,\beta}(x)+\cos\pi\lambda\tilde{k}_{\alpha,\beta}(x)=\tilde{k}_{\alpha+\lambda,\beta-\lambda}(x). \tag{2.10}$$

In particular,

$$k_{\alpha,\beta}(x)=\cos\frac{\pi\mu}{2}k_{\nu}(x)-\sin\frac{\pi\mu}{2}\tilde{k}_{\nu}(x), \tag{2.11}$$

$$\tilde{k}_{\alpha,\beta}(x)=\sin\frac{\pi\mu}{2}k_{\nu}(x)+\cos\frac{\pi\mu}{2}\tilde{k}_{\nu}(x) \tag{2.12}$$

where

$$\alpha=\tfrac{1}{2}(\mu+\nu-1), \tag{2.13}$$

$$\beta=\tfrac{1}{2}(-\mu+\nu-1). \tag{2.14}$$

Thus if we write

$$k_{\alpha,\beta}(x)=a|x|^{-\nu}+b|x|^{-\nu}\operatorname{sgn}x, \tag{2.15}$$

$$\tilde{k}_{\alpha,\beta}(x)=\tilde{a}|x|^{-\nu}+\tilde{b}|x|^{-\nu}\operatorname{sgn}x, \tag{2.16}$$

thinking of ν fixed, and given $|a|>|b|$ for a kernel of the first type and $|\tilde{a}|<|\tilde{b}|$ for a kernel of the second type, we can determine α and β from the relations

$$\frac{b}{a}=-\tan\frac{\pi\mu}{2}\tan\frac{\pi\nu}{2}, \tag{2.17}$$

$$\frac{\tilde{a}}{\tilde{b}}=\tan\frac{\pi\mu}{2}/\tan\frac{\pi\nu}{2} \tag{2.18}$$

by solving for μ and then using (2.13) and (2.14). That is, for $-1<\nu<1$ ($\nu\neq 0$) we can make $k_{\alpha,\beta}$ proportional to a given kernel of the first type by solving (2.17) for μ with $-1<\mu<1$. Since $|b/a|<1$ for a kernel of the first type we have

$$|\mu|<1-|\nu| \quad \text{for } -1<\nu<1 \text{ and } \left|\frac{b}{a}\right|<1. \tag{2.19}$$

Similarly, for $-1<\nu<1$ ($\nu\neq 0$) we can make $\tilde{k}_{\alpha,\beta}$ proportional to a given kernel of the second type by solving (2.18) for μ, $-1<\mu<1$, with $|\tilde{a}/\tilde{b}|<1$. We have

$$|\mu|<|\nu| \quad \text{for } -1<\nu<1 \text{ and } \left|\frac{\tilde{a}}{\tilde{b}}\right|<1. \tag{2.20}$$

Then from (2.13) we have

$$-1+\frac{|\nu|-\nu}{2}<\alpha<\frac{\nu-|\nu|}{2} \quad \text{under the conditions of (2.19)}, \tag{2.21}$$

$$\frac{-|\nu|+\nu-1}{2}<\alpha<\frac{|\nu|+\nu-1}{2} \quad \text{under the conditions of (2.20)}, \tag{2.21'}$$

with the same inequalities holding for β.

Thus we can find α and β in the open interval $(-1,0)$ such that $k_{\alpha,\beta}(x)$ is proportional to a given kernel of the first type with $-1<\nu<1$ ($\nu\neq 0$).

Similarly, we can find α and β in the open interval $(-\frac{1}{2},0)$ such that $\tilde{k}_{\alpha,\beta}(x)$ is proportional to a given kernel of the second type with $0<\nu<1$.

Also, whenever α and β are in the open interval $(-1,0)$, $k_{\alpha,\beta}$ is a kernel of the first type with $-1<\nu<1$, and whenever α and β are in the open interval $(-\frac{1}{2},0)$, $\tilde{k}_{\alpha,\beta}$ is a kernel of the second type with $0<\nu<1$. These statements are obvious from the definitions (2.4) and (2.5). It is easy to see that a kernel of the first or second type with any exponent can be represented as a multiple of $x^m k_{\alpha,\beta}(x)$ where m is an integer and $-1<\alpha<0$, $-1<\beta<0$.

We list here a number of useful relations which are readily derived from the basic definitions.

$$k_{\alpha,\beta}(x)=k_{\beta,\alpha}(-x), \tag{2.22}$$

$$\tilde{k}_{\alpha,\beta}(x)=\tilde{k}_{\beta,\alpha}(-x), \tag{2.23}$$

$$k_{\alpha-1/2,\beta+1/2}(x)=\tilde{k}_{\alpha,\beta}(x)=-k_{\alpha+1/2,\beta-1/2}(x), \tag{2.24}$$

$$k_{\alpha-m,\beta-n}(x)=(-1)^m x^{m+n} k_{\alpha,\beta}(x) \qquad (m,n \text{ integers}), \tag{2.25}$$

$$\tilde{k}_{\alpha-m,\beta-n}(x)=(-1)^m x^{m+n} \tilde{k}_{\alpha,\beta}(x) \qquad (m,n \text{ integers}), \tag{2.26}$$

$$\begin{aligned}\tilde{k}_{\alpha,\beta}(x)&=(-1)^m x^{m+n} k_{\alpha+m+1/2,\beta+n-1/2}(x) \qquad (m,n \text{ integers})\\ &=x k_{\alpha+1/2,\beta+1/2}(x)\\ &=-x^{-1} k_{\alpha-1/2,\beta-1/2}(x),\end{aligned} \tag{2.27}$$

$$\frac{d}{dx}k_{\alpha,\beta}(x)=-\nu\tilde{k}_{\alpha+1/2,\beta+1/2}(x)=-\nu k_{\alpha,\beta+1}(x)=-\nu x^{-1}k_{\alpha,\beta}(x), \tag{2.28}$$

$$\frac{d}{dx}\tilde{k}_{\alpha,\beta}(x)=\nu k_{\alpha+1/2,\beta+1/2}(x)=\nu\tilde{k}_{\alpha+1,\beta}(x)=-\nu x^{-1}\tilde{k}_{\alpha,\beta}(x), \tag{2.29}$$

$$\left(\frac{d}{dx}\right)^n x^m k_{\alpha,\beta}(x)=(-1)^n(\nu-m)_n x^{m-n}k_{\alpha,\beta}(x), \tag{2.30}$$

$$\left(\frac{d}{dx}\right)^n x^m \tilde{k}_{\alpha,\beta}(x)=(-1)^n(\nu-m)_n x^m m-n\tilde{k}_{\alpha,\beta}(x), \tag{2.31}$$

where m and n are integers ($n\geq 0$) and

$$(a)_n=a(a+1)\cdots(a+n-1)=(-1)^n(-a-n+1)_n$$

$$=\begin{cases}\dfrac{\Gamma(a+n)}{\Gamma(a)}, & a\neq -n,\ -n-1,\ -n-2,\cdots\\ (-1)^n\dfrac{\Gamma(1-a)}{\Gamma(-a-n+1)}, & a\neq 1,2,3\cdots,\end{cases} \tag{2.32}$$

$$(a)_0=1.$$

In the above formulas, $\nu=\alpha+\beta+1$.

3. Solution to the first problem, $-1<\nu<1$, ($\nu\neq 0$). We have the integral equation with a standard kernel of the first type

$$\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=g(x) \tag{3.1}$$

where α and β are in the open interval $(-1,0)$ and we are given only the projection of g on $(-1,1)$. We assume that f is in $L_1(-1,1)$, and with no loss in generality, we further assume that f is real valued.

We introduce the analytic function

$$\int_{-1}^{1} f(t)K_{\alpha,\beta}(z-t)\,dt=G(z). \tag{3.2}$$

G is analytic in the cut-plane where the cut removes the closed interval $[-1,1]$. Here we restrict z to the upper half plane. On the real axis we have

$$\int_{-1}^{1} f(t)K_{\alpha,\beta}(x-t)\,dt=g(x)+i\tilde{g}(x)=G(x) \tag{3.3}$$

where

$$\int_{-1}^{1} f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=\tilde{g}(x). \tag{3.4}$$

Putting $\lambda=\beta$ in (2.9) we have (cf. (2.4))

$$\cos\pi\beta k_{\alpha,\beta}(x)-\sin\pi\beta\tilde{k}_{\alpha,\beta}(x)=\begin{cases}\dfrac{1}{\pi}|x|^{-\nu}\sin\pi\nu, & x>0,\\ 0, & x<0,\end{cases} \tag{3.5}$$

$$\cos\pi\beta g(x)-\sin\pi\beta\tilde{g}(x)=\frac{1}{\pi}\sin\pi\nu\int_{-1}^{x}\frac{f(t)}{(x-t)^{\nu}}\,dt\equiv\gamma_{\nu}(x). \tag{3.6}$$

Thus if we knew the projections on $(-1,1)$ of both g and $\tilde{g}$, the problem would be solved. As we shall see, it is quite simple to determine the analytic function $G(z)$ from the projection of its real part on $(-1,1)$.

For $-1<\nu<0$, (3.6) can be differentiated with respect to x to obtain a similar equation with $0<\nu<1$. Then for $0<\nu<1$, the solution of (3.6) is simply

$$f(x)=\frac{d}{dx}\int_{-1}^{x}\frac{\gamma_\nu(t)}{(x-t)^{1-\nu}}dt, \qquad -1<x<1. \tag{3.7}$$

Now the convolution of a function of L_1 with a function of L_p, $p\geq 1$, results in a function of L_p. Although $K_{\alpha,\beta}$ does not belong to $L_p(-\infty,\infty)$ for any p, its projection on a finite interval belongs to L_p for every p satisfying $1\leq p<\nu^{-1}$ in case $0<\nu<1$, and in case $-1<\nu<0$, any projection of the derivative of $K_{\alpha,\beta}$ belongs to L_p for each p satisfying $1\leq p<(1+\nu)^{-1}$. Thus it follows that for $0<\nu<1$,

$$\int_{-T}^{T}|G(x)|^p dx<\infty \quad \text{for } 1\leq p<\nu^{-1} \text{ and } 0<T<\infty. \tag{3.8}$$

For $-1<\nu<0$, $G(x)$ is continuous and

$$\int_{-T}^{T}|G'(x)|^p dx<\infty \quad \text{for } 1\leq p<(1+\nu)^{-1} \text{ and } 0<T<\infty. \tag{3.9}$$

For large arguments, it is easy to see that

$$G(z)=K_{\alpha,\beta}(z)\int_{-1}^{1}f(t)\,dt+O\left(\frac{1}{|z|^{1+\nu}}\right), \tag{3.10}$$

as z tends to infinity on the real axis or in the upper half plane.

On the real axis, we have

$$G(x)=\begin{cases} r(x)e^{i\pi(\alpha+1/2)}, & x>1, \\ r(x)e^{-i\pi(\beta+1/2)}, & x<-1, \end{cases} \tag{3.11}$$

where $r(x)$ is a real-valued function. This follows from (2.2) and the fact that f in (3.1) is real valued.

Now consider the function $F(z)$ analytic in the upper half plane (u.h.p.) defined by

$$F(z)=W_{\alpha,\beta}(z)G(z) \tag{3.12}$$

where

$$W_{\alpha,\beta}(z)=(1-z)^{\alpha}(1+z)^{\beta} \tag{3.13}$$

and we take the branch analytic in the u.h.p. which is real in $(-1,1)$; i.e., on the real axis we have

$$W_{\alpha,\beta}(x)=\begin{cases} (1-x)^{\alpha}(1+x)^{\beta}, & -1<x<1, \\ |1-x|^{\alpha}|1+x|^{\beta}e^{-i\alpha\pi}, & x>1, \\ |1-x|^{\alpha}|1+x|^{\beta}e^{i\beta\pi}, & x<-1. \end{cases} \tag{3.14}$$

We see from (3.14) and (3.11) that

$$\operatorname{Re}F(x)=\begin{cases} 0, & |x|>1, \\ (1-x)^{\alpha}(1+x)^{\beta}g(x), & |x|<1. \end{cases} \tag{3.15}$$

Also, since

$$K_{\alpha,\beta}(z)W_{\alpha,\beta}(z)=O\left(|z|^{-1}\right), \qquad z\to\infty, \tag{3.16}$$

we have

$$F(z)=O\left(|z|^{-1}\right), \qquad z\to\infty. \tag{3.17}$$

We wish now to show that F belongs to L_p on the real axis for some $p>1$. This is sufficient (cf. Titchmarsh [13]) to obtain the representations

$$F(z)=\frac{1}{\pi i}\int_{-\infty}^{\infty}\frac{\operatorname{Re}F(t)}{t-z}dt, \qquad \operatorname{Im}z>0, \tag{3.18}$$

and

$$\operatorname{Im}F(x)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\operatorname{Re}F(t)}{x-t}dt \tag{3.19}$$

where the last integral is a Cauchy principal value at $t=x$ and is called the Hilbert transform of $\operatorname{Re}F(t)$.

In view of (3.17) it is sufficient to show that $\operatorname{Re}F(t)$ belongs to L_p for some $p>1$. Referring to (3.15), it is obvious for $-1<\nu<0$ that $\operatorname{Re}F(t)$ belongs to L_p for some $p>1$, since g is bounded and $\alpha,\beta>-1$. We are left to show that, for $0<\nu<1$,

$$\int_{-1}^{1}\left|g(t)W_{\alpha,\beta}(t)\right|^p dt<\infty \quad \text{for some } p>1. \tag{3.20}$$

We have from (3.8)

$$\int_{-1}^{1}|g(t)|^p dt<\infty \quad \text{for every } 1\le p<\nu^{-1} \quad (0<\nu<1) \tag{3.21}$$

and obviously

$$\int_{-1}^{1}\left|W_{\alpha,\beta}(t)\right|^p dt<\infty \quad \text{for every } 1<p<\rho^{-1} \tag{3.22}$$

where

$$\rho=\max(-\alpha,-\beta). \tag{3.23}$$

We have

$$\rho+\nu=\alpha+\beta+1+\max(-\alpha,-\beta)<1, \tag{3.24}$$

since $\alpha<0$, $\beta<0$.

Since $\rho+\nu<1$ it follows from Hölder's inequality that (3.20) holds for every $1\le p<(\rho+\nu)^{-1}$. To see this, set $p=r+(1-r)(\rho+\nu)^{-1}$, where $0<r\le 1$, and $s=(\rho+\nu)/\nu$ with $s'\equiv s/(s-1)=(\rho+\nu)/\rho$. Then

$$\int|g|^p|W|^p dt<\left(\int|g|^{ps}dt\right)^{1/s}\left(\int|W|^{ps'}dt\right)^{1/s'}<\infty$$

since

$$ps=\frac{r(\rho+\nu)}{\nu}+\frac{1-r}{\nu}<\frac{r}{\nu}+\frac{1-r}{\nu}=\frac{1}{\nu},$$

$$ps'=\frac{r(\rho+\nu)}{\rho}+\frac{1-r}{\rho}<\frac{r}{\rho}+\frac{1-r}{\rho}=\frac{1}{\rho}.$$

Thus (3.18) and (3.19) are valid, yielding

$$G(z)=\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi i(t-z)W_{\alpha,\beta}(z)}dt, \tag{3.25}$$

$$\tilde{g}(x)=\begin{cases}\displaystyle\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)W_{\alpha,\beta}(x)}\,dt, & -1<x<1,\\ \displaystyle\cos\alpha\pi\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)|W_{\alpha,\beta}(x)|}\,dt, & x>1,\\ \displaystyle\cos\beta\pi\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)|W_{\alpha,\beta}(x)|}\,dt, & x<-1,\end{cases}\tag{3.26}$$

$$g(x)=\begin{cases}\displaystyle-\sin\pi\alpha\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)|W_{\alpha,\beta}(x)|}\,dt, & x>1,\\ \displaystyle\sin\pi\beta\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)|W_{\alpha,\beta}(x)|}\,dt, & x<-1.\end{cases}\tag{3.27}$$

The first line in (3.26) is all that is required in (3.6). This completes the solution of the first problem for $-1<\nu<1$.

We note from (3.27) that if g is positive over $(-1,1)$, then it is positive over $(-\infty,\infty)$.

4. Solution of the second problem, $0<\nu<1$. We are given

$$\int_{-1}^{1}f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=\tilde{g}(x),\qquad -1<x<1,\tag{4.1}$$

where $\tilde{k}_{\alpha,\beta}$ is a standard kernel of the second type with α and β in the open interval $(-\frac{1}{2},0)$, and f is a real-valued function in $L_1(-1,1)$.

It follows from (2.29), the differentiation law of the kernel, that

$$\int_{-1}^{1}f(t)k_{\alpha^*,\beta^*}(x-t)\,dt=-(\alpha+\beta)\int_{0}^{x}\tilde{g}(t)\,dt+c,\qquad -1<x<1,\tag{4.2}$$

where $\alpha^*=\alpha-\frac{1}{2}$, $\beta^*=\beta-\frac{1}{2}$, and c is an arbitrary constant of integration. Now the results of the preceding section may be applied to (4.2) since k_{α^*,β^*} is necessarily a kernel of the first type with $\alpha^*+\beta^*+1=\nu^*=\nu-1$. So a separate treatment of the second problem is not necessary. However, it is instructive to see how the method used to solve the first problem must be modified to solve the second problem directly and see in another way why the solution is not unique.

Now we suppose we are given the projection on $(-1,1)$ of the imaginary part of the analytic function $G(z)$ defined by (3.2). We wish now to multiply G by a function analytic in the upper half plane which is real in $(-1,1)$ and such that the imaginary part of the product vanishes on the real line outside $[-1,1]$. Now the analytic function $F(z)$ defined by (3.12) is pure imaginary on the real axis outside $[-1,1]$, so we can either multiply or divide $F(z)$ by $(1-z^2)^{1/2}$ and obtain an analytic function whose imaginary part vanishes on the real axis outside $[-1,1]$ and the values of which we know in $(-1,1)$. However, dividing $F(z)$ by $(1-z^2)^{1/2}$ may introduce nonintegrable singularities at -1 and $+1$. We could avoid this possibility by restricting the (unknown) function f to a smaller class than $L_1(-1,1)$ (cf. Theorem 5.51). So without further restrictions on f we define the analytic function $H(z)$ by

$$H(z)=(1-z^2)^{1/2}F(z)=W_{\alpha',\beta'}(z)G(z)\tag{4.3}$$

where

$$\alpha'=\alpha+\tfrac{1}{2},\qquad \beta'=\beta+\tfrac{1}{2}.\tag{4.4}$$

We observe from (2.21) that α' and β' lie in the interval $(0,\nu)$.

We have

$$\text{Im}\,H(x)=\begin{cases}0, & |x|>1,\\ W_{\alpha',\beta'}(x)\tilde{g}(x), & |x|<1.\end{cases} \tag{4.5}$$

Now $\text{Im}\,H(x)$ belongs to L_p for $1\le p<\nu^{-1}$, since the projection of G belongs to L_p for such p and $W_{\alpha',\beta'}$ is bounded over $(-1,1)$. However, we do not have $H(z)=O(1/|z|)$, $z\to\infty$. In fact

$$H(z)=\frac{1}{\pi}\int_{-1}^{1}f(t)\,dt+O\left(\frac{1}{|z|}\right), \qquad z\to\infty. \tag{4.6}$$

So $H(z)-c$ belongs to L_p on the real axis for some $p>1$, and we may apply the boundary-value representations (3.18) and (3.19) to $iH(z)-ic$ with the results

$$G(z)=\frac{c}{W_{\alpha',\beta'}(z)}+\int_{-1}^{1}\frac{\tilde{g}(t)W_{\alpha',\beta'}(t)}{\pi(t-z)W_{\alpha',\beta'}(z)}\,dt, \tag{4.7}$$

$$g(x)=\frac{c}{W_{\alpha',\beta'}(x)}+\int_{-1}^{1}\frac{\tilde{g}(t)W_{\alpha',\beta'}(t)}{\pi(t-x)W_{\alpha',\beta'}(x)}\,dt, \qquad -1<x<1, \tag{4.8}$$

where

$$c=\frac{1}{\pi}\int_{-1}^{1}f(t)\,dt. \tag{4.9}$$

The expressions for g and $\tilde{g}$ outside $[-1,1]$ may be readily written down from (4.7). The relation (4.8) is all that is needed in (3.6) and (3.7).

We see that we can determine the projection of g only within a multiple of $(1-x)^{-\alpha'}(1+x)^{-\beta'}$ and hence can determine f only within a multiple of f_0 where f_0 is the solution of the first problem

(4.10)

$$\int_{-1}^{1}f_0(t)k_{\alpha,\beta}(x-t)\,dt=\frac{1}{\pi(1-x)^{\alpha'}(1+x)^{\beta'}}, \qquad |x|<1,\quad \alpha'=\alpha+\tfrac{1}{2},\quad \beta'=\beta+\tfrac{1}{2};$$

i.e., f_0 is a nontrivial solution to

$$\int_{-1}^{1}f_0(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=0, \qquad |x|<1. \tag{4.11}$$

From (3.6) and (3.7) we have

$$\begin{aligned}f_0(x)&=\frac{1}{\pi}\cos\pi\beta\frac{d}{dx}\int_{-1}^{x}\frac{dt}{(1-t)^{\alpha'}(1-t)^{\beta'}(x-t)^{1-\nu}}\\ &=\frac{1}{\pi}\cos\pi\beta\frac{d}{dx}\int_{0}^{1}\frac{dt}{\left(\frac{2}{1+x}-t\right)^{\alpha'}t^{\beta'}(1-t)^{1-\nu}}\\ &=\frac{2\alpha'\cos\pi\beta}{\pi(1+x)^2}\int_{0}^{1}\frac{dt}{\left(\frac{2}{1+x}-t\right)^{\alpha'+1}t^{\beta'}(1-t)^{1-\nu}}.\end{aligned} \tag{4.12}$$

If in the last line of (4.12) we put $t=u((1-x)/2+u)^{-1}$ with $0<u<\infty$, we obtain

$$f_0(x)=\frac{2^{1-\nu}\alpha'\cos\pi\beta}{\pi(1-x)^{1-\beta'}(1+x)^{1-\alpha'}}\int_0^\infty \frac{du}{(u+1)^{1+\alpha'}u^{\beta'}}$$

$$=\frac{2^{1-\nu}\Gamma(\nu)}{\Gamma(\alpha')\Gamma(\beta')}(1-x)^{\beta'-1}(1+x)^{\alpha'-1}, \qquad |x|<1, \tag{4.13}$$

and

$$\int_{-1}^{1} f_0(t)\,dt=1, \tag{4.14}$$

which is in accord with (4.8) and (4.9).

We can also deduce the form of f_0 by considering the function analytic in the upper half plane

$$A(1-z)^a(1-z)^b(z-\tau)^c, \qquad a+b+c=-2, \quad a,b,c>-1, \tag{4.15}$$

where $-1<\tau<1$. The function belongs to L_1 on the real axis, implying, with the analyticity, that its integral over $(-\infty,\infty)$ vanishes. With the proper choice of A we obtain a relation of the form (4.11). We obtain the same result a different way in the following section.

First let us note the following generalization of the previous result.

We see from (4.7) and (4.13) that

$$\frac{2^{1-\nu}\Gamma(\nu)}{\Gamma(\lambda)\Gamma(\mu)}\int_{-1}^{1}(1-t)^{\mu-1}(1+t)^{\lambda-1}K_{\lambda',\mu'}(x-t)\,dt=\frac{1}{\pi}W_{-\lambda,-\mu}(x), \tag{4.16}$$

$$-\infty<x<\infty,$$

where $\mu>0$, $\lambda>0$, $\lambda+\mu=\nu<1$, $\lambda'=\lambda-\frac{1}{2}$, $\mu'=\mu-\frac{1}{2}$.

Here $K_{(\cdot,\cdot)}$ is the analytic kernel defined in (2.2) and $W_{(\cdot,\cdot)}$ is defined in (3.14). We have

$$K_{\alpha,\beta}(x)=e^{i\pi(\alpha-\lambda')}K_{\lambda',\mu'}(x). \tag{4.17}$$

Thus for $\mu+\lambda=\alpha+\beta+1=\nu<1$ and $\mu>0$, $\lambda>0$ we have

$$\frac{2^{1-\nu}\Gamma(\nu)}{\Gamma(\lambda)\Gamma(\mu)}\int_{-1}^{1}(1-t)^{\mu-1}(1+t)^{\lambda-1}k_{\alpha,\beta}(x-t)\,dt$$

$$=\begin{cases}\frac{1}{\pi}\{\sin\pi(\lambda-\alpha)\}(1-x)^{-\lambda}(1+x)^{-\mu}, & -1<x<1,\\ -\frac{1}{\pi}\{\sin\pi\alpha\}|1-x|^{-\lambda}|1+x|^{-\mu}, & x>1,\\ -\frac{1}{\pi}\{\sin\pi\beta\}|1-x|^{-\lambda}|1+x|^{-\mu}, & x<-1;\end{cases} \tag{4.18}$$

$$\frac{2^{1-\nu}\Gamma(\nu)}{\Gamma(\lambda)\Gamma(\mu)}\int_{-1}^{1}(1-t)^{\mu-1}(1+t)^{\lambda-1}\tilde{k}_{\alpha,\beta}(x-t)\,dt$$

$$=\begin{cases}\frac{1}{\pi}\{\cos\pi(\lambda-\alpha)\}(1-x)^{-\lambda}(1-x)^{-\mu}, & -1<x<1,\\ \frac{1}{\pi}\{\cos\pi\alpha\}|1-x|^{-\lambda}|1+x|^{-\mu}, & x>1,\\ -\frac{1}{\pi}\{\cos\pi\beta\}|1-x|^{-\lambda}|1+x|^{-\mu}, & x<-1.\end{cases} \tag{4.19}$$

5. The connection with the Jacobi polynomials. In the first or second problem, if the integral agrees over $(-1,1)$ with a polynomial, then the solution, if it exists, is an elementary function. The solution always exists if the kernel is of the first type with $-1<\nu<1$ $(\nu\neq 0)$ or if the kernel is of the second type with $0<\nu<1$. In the latter case the solution is not unique, as we have seen. In the general case, there are constraints on the polynomials in order for the problem to have a solution. These results we derive by considering the asymptotic behavior of $g(x)$ as $x\to\infty$ where

$$\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=g(x) \tag{5.1}$$

and $-1<\alpha<0$, $-1<\beta<0$, i.e., $k_{\alpha,\beta}$ is a kernel of the first type with $-1<\nu<1$.

From (2.4) and (3.27) we have

$$\operatorname{sgn}x\int_{-1}^{1}\frac{f(t)\,dt}{|x-t|^{\nu}}=|x-1|^{-\alpha}|x+1|^{-\beta}\int_{-1}^{1}\frac{W_{\alpha,\beta}(t)g(t)}{x-t}\,dt, \qquad |x|>1. \tag{5.2}$$

Then by replacing x by x^{-1} and using the fact that $\alpha+\beta+1=\nu$, we obtain

$$(1-x)^{\alpha}(1+x)^{\beta}\int_{-1}^{1}\frac{f(t)\,dt}{(1-xt)^{\nu}}=\int_{-1}^{1}\frac{W_{\alpha,\beta}(t)g(t)}{1-xt}\,dt, \qquad -1<x<1. \tag{5.3}$$

(It is worth noting that we know how to solve (5.3) for f or g, given the other.)

Equating coefficients of x^n in (5.3) we obtain

$$\int_{-1}^{1} W_{\alpha,\beta}(t)t^{n}g(t)\,dt=\int_{-1}^{1} p_n(t)f(t)\,dt \tag{5.4}$$

where $p_n(t)$ is a polynomial of degree n defined by

$$\frac{(1-x)^{\alpha}(1+x)^{\beta}}{(1-xt)^{\nu}}=\sum_{0}^{\infty}p_n(t)x^n, \qquad -1<x<1, \quad -1<t<1 \quad (\nu=\alpha+\beta+1). \tag{5.5}$$

Similarly, if we divide both sides of (5.3) by $(1-x)^{\alpha}(1+x)^{\beta}$ and then equate coefficients of x^n, we obtain

$$\int_{-1}^{1} f(t)\frac{(\nu)_n}{n!}t^{n}\,dt=\int_{-1}^{1} W_{\alpha,\beta}(t)q_n(t)g(t)\,dt \tag{5.6}$$

where $q_n(t)$ is a polynomial of degree n defined by

$$\frac{(1-x)^{-\alpha}(1+x)^{-\beta}}{1-xt}=\sum_{0}^{\infty}q_n(t)x^n, \qquad -1<x<1, \quad -1<t<1, \tag{5.7}$$

and $(\nu)_n$ is defined in (5.21).

Then to each polynomial $P_n(t)$ there corresponds a polynomial $P_n^*(t)$ of the same degree such that

$$\int_{-1}^{1} W_{\alpha,\beta}(t)P_n(t)g(t)\,dt=\int_{-1}^{1} P_n^*(t)f(t)\,dt. \tag{5.8}$$

This relation holds for every integrable f and its transform g under the convolution kernel $k_{\alpha,\beta}$; i.e., for arbitrary f in L_1,

$$\int_{-1}^{1} W_{\alpha,\beta}(x)P_n(x)\,dx\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=\int_{-1}^{1} f(t)P_n^*(t)\,dt. \tag{5.9}$$

We wish to change the order of integration in (5.9). This can be justified by the Tonelli–Hobson theorem, which says that if either of the iterated integrals

$$I_1=\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}\varphi(x,t)\,dt,\qquad I_2=\int_{-\infty}^{\infty}dt\int_{-\infty}^{\infty}\varphi(x,t)\,dx$$

is absolutely convergent, then both are, and

$$I_1=I_2=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\varphi(x,t)\,dx\,dt,$$

and the double integral is absolutely convergent. In particular, suppose $f\in L_1$, $k\in L_p$ for some p $(1\le p\le\infty)$, and

$$\int_{-\infty}^{\infty}f(t)k(x-t)\,dt=g(x). \tag{5.10}$$

Then for h in L_q where $\frac{1}{p}+\frac{1}{q}=1$ we have

$$\int_{-\infty}^{\infty}h(t)g(t)\,dt=\int_{-\infty}^{\infty}f(t)\hat{h}(t)\,dt \tag{5.11}$$

where

$$\hat{h}(t)=\int_{-\infty}^{\infty}h(x)k(x-t)\,dx, \tag{5.12}$$

for we have

$$\int_{-\infty}^{\infty}|f(t)|\,dt\int_{-\infty}^{\infty}|h(x)|\,|k(x-t)|\,dx\le\|f\|_1\|h\|_q\|k\|_p.$$

So the order of integration may be reversed, i.e.,

$$\int_{-\infty}^{\infty}h(x)\,dx\int_{-\infty}^{\infty}f(t)k(x-t)\,dt=\int_{-\infty}^{\infty}f(t)\,dt\int_{-\infty}^{\infty}h(x)k(x-t)\,dx. \tag{5.13}$$

Now we let $k(x)$ be the projection on $(-2,2)$ of $k_{\alpha,\beta}(x)$ and let $h(x)$ be the projection on $(-1,1)$ of $W_{\alpha,\beta}(x)P_n(x)$. Then for $-1<\nu<0$, k is bounded and h is in L_1. For $0<\nu<1$, k belongs to L_p where $p=(\rho+\nu)/\nu$ and h belongs to L_q where $q=(\rho+\nu)/\rho$ with $\rho=\max(-\alpha,-\beta)$ (cf. (3.24)). Thus if we integrate first on x in (5.9), the iterated integral is absolutely convergent, so the change in order of integration is justified. Thus

$$\int_{-1}^{1}W_{\alpha,\beta}(x)P_n(x)g(x)\,dx=\int_{-1}^{1}f(t)\hat{P}_n(t)\,dt=\int_{-1}^{1}f(t)P_n^*(t)\,dt \tag{5.14}$$

where

$$\hat{P}_n(t)=\int_{-1}^{1}W_{\alpha,\beta}(x)P_n(x)k_{\alpha,\beta}(x-t)\,dx. \tag{5.15}$$

Since (5.14) must hold for arbitrary f in L_1, we have

$$\hat{P}_n(t)\equiv P_n^*(t),\qquad |t|<1. \tag{5.16}$$

Replacing x by $-x$ and t by $-t$ in (5.15) and noting that $W_{\alpha,\beta}(-x)=W_{\beta,\alpha}(x)$, we have

$$\int_{-1}^{1}W_{\beta,\alpha}(x)P_n(-x)k_{\alpha,\beta}(t-x)\,dx=P_n^*(-t),\qquad -1<t<1. \tag{5.17}$$

With this special pair of f and g in (5.3) we have

$$(1-x)^{\alpha}(1+x)^{\beta}\int_{-1}^{1}\frac{W_{\beta,\alpha}(t)P_n(-t)\,dt}{(1-xt)^{\nu}}=\int_{-1}^{1}\frac{W_{\alpha,\beta}(t)P_n^*(-t)\,dt}{1-xt}, \tag{5.18}$$

$$-1<x<1.$$

Now if we set $P_n(-t)=P_n^{(\beta,\alpha)}(t)$, the Jacobi polynomial of degree n associated with the weight function $W_{\beta,\alpha}(t)$, the left side of (5.18) is $O(x^n)$ as $x\to 0$. Therefore, for that choice of P_n, we must have $P_n^*(-t)=\lambda_n P_n^{(\alpha,\beta)}(t)$. Thus

$$\int_{-1}^{1} W_{\beta,\alpha}(t)P_n^{(\beta,\alpha)}(t)k_{\alpha,\beta}(x-t)\,dt=\lambda_n P_n^{(\alpha,\beta)}(x), \tag{5.19}$$

$$|x|<1 \quad (-1<\alpha<0) \quad (-1<\beta<0).$$

Equating powers of x^n in (5.18) we have

$$\int_{-1}^{1} W_{\beta,\alpha}(t)P_n^{(\beta,\alpha)}(t)\frac{(\nu)_n}{n!}t^n\,dt=\lambda_n\int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)t^n\,dt \tag{5.20}$$

where

$$(\nu)_n=\nu(\nu+1)\cdots(\nu+n-1)=\frac{\Gamma(\nu+n)}{\Gamma(\nu)}, \qquad \nu=\alpha+\beta+1. \tag{5.21}$$

We follow the notation and standardization employed by Szegö [12]. We have

$$P_n^{(\alpha,\beta)}(1)=\binom{n+\alpha}{n}=\frac{\Gamma(n+\alpha+1)}{n!\Gamma(\alpha+1)}, \tag{5.22}$$

$$P_n^{(\alpha,\beta)}(x)=(-1)^n P_n^{(\beta,\alpha)}(-x). \tag{5.23}$$

Using (5.23) and (5.20) we find

$$\lambda_n=\frac{(\nu)_n}{n!}. \tag{5.24}$$

So with f and g related by (5.1) we have

$$\int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)g(t)\,dt=\frac{(\nu)_n}{n!}\int_{-1}^{1} f(t)P_n^{(\beta,\alpha)}(t)\,dt \tag{5.25}$$

$$(-1<\alpha<0) \quad (-1<\beta<0).$$

We may expand g in a series of Jacobi polynomials and get a series representation of the solution of (4.1); viz., if

$$g(x)\sim\sum_{0}^{\infty} a_n P_n^{(\alpha,\beta)}(x), \qquad -1<x<1, \tag{5.26}$$

then

$$f(x)\sim\sum_{0}^{\infty}\frac{n!a_n}{(\nu)_n}P_n^{(\beta,\alpha)}(x)W_{\beta,\alpha}(x), \qquad -1<x<1, \tag{5.27}$$

where

$$a_n=h_n^{-1}\int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)g(t)\,dt \tag{5.28}$$

and

$$\begin{aligned} h_n\equiv h_n(\alpha,\beta)=h_n(\beta,\alpha)&=\int_{-1}^{1}\{P_n^{(\alpha,\beta)}(t)\}^2 W_{\alpha,\beta}(t)\,dt \\ &=\frac{2^\nu}{2n+\nu}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\nu)}. \end{aligned} \tag{5.29}$$

We pause here to record some results which follow from (5.4)–(5.7), and (5.17), namely

$$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}t^{n}k_{\alpha,\beta}(x-t)\,dt=A_n^{(\alpha,\beta)}(x), \tag{5.30}$$

$$-1<x<1,\quad \alpha>-1,\quad \beta>-1,\quad \alpha+\beta<0$$

where $A_n^{(\alpha,\beta)}$ is a polynomial of degree n defined by

$$\frac{(1+x)^{\alpha}(1-x)^{\beta}}{(1-xt)^{\nu}}=\sum_{0}^{\infty}A_n^{(\alpha,\beta)}(t)x^{n},\qquad \nu=\alpha+\beta+1. \tag{5.31}$$

We have

$$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}B_n^{(\beta,\alpha)}(t)k_{\alpha,\beta}(x-t)\,dt=\frac{(\nu)_n}{n!}x^{n}, \tag{5.32}$$

$$-1<x<1,\quad \alpha>-1,\quad \beta>-1,\quad \nu=\alpha+\beta+1<1,$$

where $B_n^{(\beta,\alpha)}$ is a polynomial of degree n defined by

$$\frac{(1-x)^{-\beta}(1+x)^{-\alpha}}{1-xt}=\sum_{0}^{\infty}B_n^{(\beta,\alpha)}(t)x^{n}. \tag{5.33}$$

Note that the leading coefficient of $A_n^{(\alpha,\beta)}$ is $(\nu)_n/n!$ and that of $B_n^{(\beta,\alpha)}$ is unity.

Also note that if we multiply both sides of (5.3) by $x^nP_n^{(\alpha,\beta)}(\frac{1}{x})$ and equate coefficients of x^n we obtain

$$\int_{-1}^{1}C_{n,n}^{(\beta,\alpha)}(t)f(t)\,dt=\int_{-1}^{1}P_n^{(\alpha,\beta)}(t)W_{\alpha,\beta}(t)g(t)\,dt \tag{5.34}$$

where $C_{n,k}^{(\beta,\alpha)}$ is a polynomial of degree k defined by

$$x^{n}P_n^{(\alpha,\beta)}(\tfrac{1}{x})\frac{(1-x)^{\alpha}(1+x)^{\beta}}{(1-xt)^{\nu}}=\sum_{k=0}^{\infty}C_{n,k}^{(\beta,\alpha)}(t)x^{k},\qquad \nu=\alpha+\beta+1. \tag{5.35}$$

It follows from (5.25) that

$$C_{n,n}^{(\beta,\alpha)}(t)=\frac{(\nu)_n}{n!}P_n^{(\beta,\alpha)}(t). \tag{5.36}$$

Returning to (5.19), if $\nu<0$ (i.e., $\alpha+\beta<-1$), the equation may be differentiated with respect to x, using (2.28) and

$$\left(\frac{d}{dx}\right)^{m}P_n^{(\alpha,\beta)}(x)=\left(\frac{1}{2}\right)^{m}(n+\alpha+\beta+1)_mP_{n-m}^{(\alpha+m,\beta+m)}(x), \tag{5.37}$$

to obtain

$$\int_{-1}^{1}(1-t)^{\beta-1/2}(1+t)^{\alpha-1/2}P_n^{(\beta-1/2,\alpha-1/2)}(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt$$

$$=-\frac{1}{2}\frac{(\nu)_n}{n!}P_{n-1}^{(\alpha+1/2,\beta+1/2)}(x), \tag{5.38}$$

$$(-1<x<1),\quad \nu=\alpha+\beta+1<1,\quad -\frac{1}{2}<\alpha,\quad -\frac{1}{2}<\beta,$$

where $P_{-1}^{(\cdot)}\equiv 0$.

We would expect (5.19) to hold over the triangle $\{-1<\alpha,\ -1<\beta,\ \alpha+\beta<0\}$ since the integral still makes sense over this extended region. However, it is only over the

square $\{-1<\alpha<0,\ -1<\beta<0\}$ that (the projections of) $W_{\beta,\alpha}$ and $k_{\alpha,\beta}$ belong to complementary L_p spaces for certain $p>1$, which fact was used to establish the validity of (5.25) as well as (5.19) itself. In fact, (5.25) is not generally valid over the triangle as we can see by rewriting (5.38) as

$$\int_{-1}^{1}(1-t)^{\beta-1}(1+t)^{\alpha}P_n^{(\beta-1,\alpha)}(t)k_{\alpha,\beta}(x-t)\,dt=-\frac{1}{2}\frac{(\nu)_n}{n!}P_{n-1}^{(\alpha+1,\beta)}(x), \tag{5.39}$$

$$-1<x<1,\quad \alpha>-1,\quad \beta>0,\quad \alpha+\beta<0\quad (\nu=\alpha+\beta+1).$$

Here we have made use of (2.24). Over the triangle conditioned in (5.39), $k_{\alpha,\beta}$ is a kernel of the second type with $0<\nu<1$, and for $n=0$ in (5.39) we have an (f,g) pair for which (5.25) fails for $n=0$ and $\beta>0$. On the other hand, (5.19) is valid over the larger region. In the Appendix we derive the more general result

$$\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}P_n^{(\beta+r,\alpha+s)}(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt \tag{5.40}$$

$$=\frac{2^{m+r+s}(\alpha+\beta+1-m)_n}{(-1)^{m+r}n!}P_{n^*}^{(\alpha^*,\beta^*)}(x),\qquad -1<x<1,$$

where r,s,m, and n are integers $(n\geq 0)$, $\beta+r>-1$, $\alpha+s>-1$, $m-\alpha-\beta-1>-1$, $\alpha^*=\alpha-m-r$, $\beta^*=\beta-m-s$, $n^*=m+r+s+n\geq -1$ and $P_{-1}^{(\cdot)}\equiv 0$.

It is no restriction to assume that $-1<\alpha<0$ and $-1<\beta<0$ in (5.40). The result (5.39), for example, can then be obtained by putting $r=s=0$ and $m=-1$, using the relation (2.25),

$$x^{-1}k_{\alpha,\beta}(x)=k_{\alpha,\beta+1}(x).$$

Incidentally, if we divide both sides of (5.40) by $\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}dt$ and set $r=s$, $n=0$, we obtain, letting $r\to\infty$,

$$\lim_{r\to\infty}2^{m-\nu}(-1)^{m+r}\frac{\Gamma(\beta+\alpha+2+2r)}{\Gamma(\beta+r+1)\Gamma(\alpha+r+1)}P_{2r+m}^{(\alpha-m-r,\beta-m-r)}(x)=x^m k_{\alpha,\beta}(x), \tag{5.41}$$

$$-1<x<1,\quad m-(\alpha+\beta)>0,\quad \nu=\alpha+\beta+1,\quad m \text{ and } r \text{ integers}.$$

Using $\tilde{k}_{\alpha,\beta}=k_{\alpha-1/2,\beta-1/2}$, we have as a special case of (5.40) with $r=0$, $s=1$, $m=0$,

(5.42)

$$\int_{-1}^{1}(1-t)^{\beta+1/2}(1+t)^{\alpha+1/2}P_n^{(\beta+1/2,\alpha+1/2)}(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=2\frac{(\nu)_n}{n!}P_{n+1}^{(\alpha-1/2,\beta-1/2)}(x),$$

$$-1<x<1,\quad \alpha>-\frac{3}{2},\quad \beta>-\frac{3}{2},\quad \nu=\alpha+\beta+1<1,\quad n=0,1,2,\cdots.$$

Now if f in L_1 and $\tilde{g}$ are related by

$$\int_{-1}^{1}f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=\tilde{g}(x), \tag{5.43}$$

$$-\frac{3}{2}<\alpha<\frac{1}{2},\quad -\frac{3}{2}<\beta<\frac{1}{2},\quad \alpha+\beta<0,$$

we have from (5.42), using $\tilde{k}_{\alpha,\beta}(-x)=-\tilde{k}_{\beta,\alpha}(x)$,

$$\int_{-1}^{1}(1-t)^{\alpha+1/2}(1+t)^{\beta+1/2}P_n^{(\alpha+1/2,\beta+1/2)}(t)\tilde{g}(t)\,dt \tag{5.44}$$

$$=-2\frac{(\nu)_n}{n!}\int_{-1}^{1}P_{n+1}^{(\beta-1/2,\alpha-1/2)}(t)f(t)\,dt,\qquad \nu=\alpha+\beta+1,\quad n=0,1,2,\cdots,$$

over the region in the $\alpha-\beta$ plane conditioned in (5.43). The projection of $W_{\alpha+1/2,\beta+1/2}$ and $\tilde{k}_{\alpha,\beta}$ belong to complementary L_p spaces, and the interchange of order of integration leading to (5.44) is justified. It should be noted that this region includes the region over which $\tilde{k}_{\alpha,\beta}$ is a kernel of the second type with $-2<\nu<1$, viz., over the square $\{-\frac{3}{2}<\alpha<-\frac{1}{2},\ -\frac{3}{2}<\beta<-\frac{1}{2}\}$ and the triangle $\{\alpha>-\frac{1}{2},\ \beta>-\frac{1}{2},\ \alpha+\beta<0\}$.

We can see also from (5.44) why, in the second problem for $0<\nu<1$, we cannot in general determine $\int_{-1}^{1} f\,dt$ from the projection of $\tilde{g}$. However, if we restrict f to a suitable class, we can. Suppose then that

$$\int_{-1}^{1} f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=\tilde{g}(x) \tag{5.45}$$

where $\alpha>-\frac{1}{2}$, $\beta>-\frac{1}{2}$, $\alpha+\beta<0$. Now suppose $\varphi(x)$ is a bounded function and we consider

$$\begin{aligned}&\int_{-1}^{1}\varphi(x)(1-x)^{\alpha-1/2}(1+x)^{\beta-1/2}\tilde{g}(x)\,dx\\ &\qquad=\int_{-1}^{1}\varphi(x)(1-x)^{\alpha-1/2}(1+x)^{\beta-1/2}\left\{\int_{-1}^{1} f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt\right\}dx,\end{aligned} \tag{5.46}$$

assuming the integral is convergent. Denoting the integral by I, we have

$$|I|\le c\int_{-1}^{1}(1-x)^{\alpha-1/2}(1+x)^{\beta-1/2}\left\{\int_{-1}^{1}|f(t)|\,|\tilde{k}_{\alpha,\beta}(x-t)|\,dt\right\}dx \tag{5.47}$$

and

$$\left|\tilde{k}_{\alpha,\beta}(x)\right|\le\frac{1}{\pi}|x|^{-\nu}=\sec\frac{\pi\nu}{2}k_{\nu}(x) \tag{5.48}$$

and from (2.29)

$$k_{\nu}(x)=\cos\pi\frac{(\beta-\alpha)}{2}k_{\alpha,\beta}(x)-\sin\pi\frac{(\beta-\alpha)}{2}\tilde{k}_{\alpha,\beta}(x). \tag{5.49}$$

Now assuming that f is suitably restricted so that the order of integration can be changed in (5.46), we find from (4.18) and (4.19)

$$|I|\le c'\int_{-1}^{1}|f(t)|(1-t)^{-\beta'}(1+t)^{-\alpha'}dt \tag{5.50}$$

where $\alpha'=\alpha+\frac{1}{2}$, $\beta'=\beta+\frac{1}{2}$. So if the last integral is convergent, the change of the order of integration is justified by the Tonelli–Hobson theorem, and we have the result:

THEOREM 5.51. *Let $L^{(\mu,\lambda)}$ be the class of functions $\{f\}$ satisfying*

$$\int_{-1}^{1}|f(t)|(1-t)^{-\mu}(1+t)^{-\lambda}dt<\infty \tag{5.52}$$

where λ and μ are real numbers.

Then if f in (5.45) *belongs to $L^{(\beta',\alpha')}$ where $\beta'=\beta+\frac{1}{2}$, $\alpha'=\alpha+\frac{1}{2}$, we have*

(5.53)

$$\int_{-1}^{1}(1-x)^{\alpha-1/2}(1+x)^{\beta-1/2}P_n^{(\alpha-1/2,\beta-1/2)}\tilde{g}(x)\,dx=\frac{(\nu)_n}{2(n!)}\int_{-1}^{1}P_{n-1}^{(\beta',\alpha')}(t)f(t)\,dt$$

where $\nu=\alpha+\beta+1$ and $n=0,1,2,\cdots,P_{-1}^{(\cdot)}\equiv 0$. In particular,

$$\int_{-1}^{1}(1-x)^{\alpha-1/2}(1+x)^{\beta-1/2}\tilde{g}(x)\,dx=0. \tag{5.54}$$

On the other hand, for given $\tilde{g}$, if (5.45) *has a solution f in $L^{(\beta',\alpha')}$ it is unique and* (5.54) *is a necessary condition for such a solution.*

(The relation (5.53) follows from (5.38), using $\tilde{k}_{\alpha,\beta}(-x)=-\tilde{k}_{\beta,\alpha}(x)$ and the assumption $f\in L^{(\beta',\alpha')}$ which allows the change of order of integration. Also the assumption on f excludes the nontrivial solution f_0 given by (4.13) for the homogeneous equation.)

Now if f and g satisfy

$$\int_{-1}^{1} f(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt=g(x), \tag{5.55}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0,$$

for some nonnegative integer m, it follows from (5.40) that

$$\int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)g(t)\,dt=2^m\frac{(\nu-m)_n}{n!}\int_{-1}^{1} P_{n+m}^{(\beta^*,\alpha^*)}(t)f(t)\,dt \tag{5.56}$$

where $\alpha^*=\alpha-m$, $\beta^*=\beta-m$, $\nu=\alpha+\beta+1$. Incidentally, we have a small theorem in (5.56):

THEOREM 5.57. *If f belongs to $L_1(-1,1)$ and*

$$\int_{-1}^{1} f(t)P_n^{(\alpha,\beta)}(t)\,dt=0 \quad \text{for } n=m, m+1, m+2, \cdots$$

where m is a nonnegative integer and $-(m+1)<\alpha<-m$, $-(m+1)<\beta<-m$, then f is null over $(-1,1)$.

For the proof, we observe that the left-hand side of (5.56) can vanish for $n=0,1,2,\cdots$ only if g is null. But from the relation (5.55) and the solution to the first problem, g can be null only if f is null.

The interchange of the order of integration leading to (5.56) is justified as before (we have excluded $m=-1$).

In solving (5.55), the equation is differentiated m times to obtain

$$(-1)^m(\nu-m)_m\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=\left(\frac{d}{dx}\right)^m g(x), \tag{5.58}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0.$$

Now a polynomial of degree $m-1$ can be added to g without affecting the solution of (5.58), but $g^{(m)}(x)$ determines f and hence g through (5.55). In other words,

$$g(x)=\idotsint_{m}^{x} g^{(m)}(t)\,dt+\pi_{m-1}(x),$$

where π_{m-1} is a polynomial of degree $m-1$ which is determined by the relation (5.56) for $n=0,1,2,m-1$. We can conclude:

THEOREM 5.59. *If f is a solution of* (5.58), *it is also a solution of* (5.55) *if and only if* (5.56) *is satisfied for $n=0,1,2,\cdots m-1$. Furthermore, if g agrees over $(-1,1)$ with a polynomial of degree n, then* (5.55) *has a solution if and only if*

$$g(x)=\sum_{k=0}^{n-m} a_k P_{k+m}^{(\alpha^*,\beta^*)}(x), \qquad -1<x<1, \tag{5.60}$$

where $\alpha^*=\alpha-m$, $\beta^*=\beta-m$, $n\geq m$.

Proof. From (5.58), $f(t)=\Sigma_0^{n-m}b_k P_k^{(\beta,\alpha)}(t)W_{\beta,\alpha}(t)$, $-1<t<1$. Then (5.60) follows from (5.40). □

There is an equivalent statement of the first part of Theorem 5.59. Actually, the explicit solution of (5.58) is not needed in the test since (5.58) implies

$$(5.61)\quad \frac{2^m(\nu-m)_n}{n!}\int_{-1}^{1} P_{n+m}^{(\beta^*,\alpha^*)}f(t)\,dt=\frac{(-1)^m}{(\nu-m)_m}\int_{-1}^{1} W_{\alpha,\beta}(t)G_{m+n}^{(\alpha,\beta)}(t;m)g^{(m)}(t)\,dt$$

where $g^{(m)}(t)=(\frac{d}{dt})^m g(t)$, and $G_{m+n}^{(\alpha,\beta)}$ is a polynomial of degree $m+n$ determined by

$$(5.62)\quad \int_{-1}^{1} W_{\beta,\alpha}(t)G_{m+n}^{(\beta,\alpha)}(t;m)k_{\alpha,\beta}(x-t)\,dt=2^m\frac{(\nu-m)_n}{n!}P_{n+m}^{(\alpha^*,\beta^*)}(x),$$

with $-1<x<1$, $-1<\alpha<0$, $-1<\beta<0$, $\nu=\alpha+\beta+1<1$, $\alpha^*=\alpha-m$, $\beta^*=\beta-m$, and $m=$ positive integer. Thus we can state

THEOREM 5.63. *If f is a solution of* (5.58), *then f is also a solution of* (5.55) *if and only if*

$$(5.64)\quad (1-\nu)_m\int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)g(t)\,dt=\int_{-1}^{1} W_{\alpha,\beta}(t)G_{m+n}^{(\alpha,\beta)}(t;m)g^{(m)}(t)\,dt$$

for $n=0,1,2,\cdots m-1$ where $G_{m+n}^{\alpha,\beta}$ is determined by (5.62). (*Here we have used the identity $(-1)^m(\nu-m)_m=(1-\nu)_m$.*)

We note that (5.64) is satisfied for $n\geq m$ for any function g which is m times differentiable. We have from (5.40), for $n\geq m$,

(5.65)

$$\int_{-1}^{1} W_{\beta,\alpha}(t)(1-t^2)^m P_{n-m}^{(\beta+m,\alpha+m)}(t)k_{\alpha,\beta}(x-t)\,dt=(-4)^m\frac{(\nu)_{n-m}}{(n-m)!}P_{n+m}^{(\alpha^*,\beta^*)}(x),$$

$$-1<x<1,\quad \alpha>-1,\quad \beta>-1,\quad \nu=\alpha+\beta+1<1,\quad \alpha^*=\alpha-m,\quad \beta^*=\beta-m.$$

So we see that

$$(5.66)\quad G_{n+m}^{(\alpha,\beta)}(t;m)=\left(-\frac{1}{2}\right)^m\frac{(n-m)!}{n!}(\nu-m)_m(1-t^2)^m P_{n-m}^{(\alpha+m,\beta+m)}(t),\qquad n\geq m.$$

Also, for $n\geq m$ we have

$$(5.67)\quad W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)=\left(-\frac{1}{2}\right)^m\frac{(n-m)!}{n!}\left(\frac{d}{dt}\right)^m\left\{W_{\alpha+m,\beta+m}(t)P_{n-m}^{(\alpha+m,\beta+m)}\right\},$$

which is a generalization of Rodrigues' formula ($n=m$). Then if g is any function m times differentiable we have from (5.67)

$$(5.68)\quad \int_{-1}^{1} W_{\alpha,\beta}(t)P_n^{(\alpha,\beta)}(t)g(t)\,dt$$
$$=\left(\frac{1}{2}\right)^m\frac{(n-m)!}{n!}\int_{-1}^{1} W_{\alpha+m,\beta+m}(t)P_{n-m}^{(\alpha+m,\beta+m)}(t)g^{(m)}(t)\,dt,$$
$$n\geq m,\quad \alpha>-1,\quad \beta>-1.$$

i.e., (5.64) has nothing to do with the integral equation for $n\geq m$.

There is still another equivalent statement of Theorem 5.59. From (5.55) we have

$$(5.69)\quad (-1)^k(\nu-m)_k\int_{-1}^{1} f(t)(x-t)^{m-k}k_{\alpha,\beta}(x-t)\,dt=\left(\frac{d}{dx}\right)^k g(x)\equiv g^{(k)}(x),$$
$$-1<\alpha<0,\quad -1<\beta<0,\quad k=0,1,2,\cdots m.$$

Since

$$g^{(k-1)}(x)=\int^x g^{(k)}(t)\,dt+c_k \tag{5.70}$$

and

$$g(x)=\iiint_m^x g^{(m)}(t)\,dt+\pi_{m-1}(x), \tag{5.71}$$

the determination of the m constants c_k, $k=1,2,\cdots m$, is equivalent to determining the polynomial π_{m-1} by use of (5.64) for $n=0,1,2,\cdots m-1$. The constants c_k can be determined by applying (5.64) with $n=0$ to (5.69). We have

$$(1-\nu)_k\int_{-1}^1 W_{\alpha,\beta}(t)g^{(m-k)}(t)\,dt=\int_{-1}^1 W_{\alpha,\beta}(t)G_k^{(\alpha,\beta)}(t;k)g^{(m)}(t)\,dt, \tag{5.72}$$

$$k=1,2,3,\cdots m,$$

where $g^{(k)}$ is given by (5.69) and $\nu=\alpha+\beta+1$. Thus we have:

THEOREM 5.73. *The condition* (5.64) *in Theorem* 5.63 *may be replaced by the equivalent condition* (5.72).

Explicit expressions for the G polynomials are rather messy for cases other than given in (5.66). They may be found by expanding the right-hand member of (5.62) as $\Sigma a_k((1-x)/2)^k$ and using a later result (6.52). Alternatively, the right-hand member may be expanded as $\Sigma b_k P_k^{(\alpha,\beta)}$ by repeated application of the formula

$$\begin{aligned} P_n^{(\alpha-1,\beta-1)}(x)=&\frac{(n+\alpha+\beta)(n+\alpha+\beta-1)}{(2n+\alpha+\beta)(2n+\alpha+\beta-1)}P_n^{(\alpha,\beta)}(x)\\ &+\frac{(n+\alpha+\beta-1)(\alpha-\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta-2)}P_{n-1}^{(\alpha,\beta)}(x)\\ &-\frac{(n+\beta-1)(n+\alpha-1)}{(2n+\alpha+\beta-1)(2n+\alpha+\beta-2)}P_{n-2}^{(\alpha,\beta)}(x) \end{aligned} \tag{5.74}$$

and then using the basic result (5.19) to solve (5.62).

The perplexing thing about (5.72) is the fact that the integral on the right cannot be replaced by an integral involving only $g^{(m-k+1)}$ (except for $k=1$). To see this, suppose

$$\int_{-1}^1 f(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt=g_m(x) \tag{5.75}$$

and

$$\int_{-1}^1 f(t)(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dt=g_{m-1}(x) \tag{5.76}$$

where m is an integer, $m\geq 2$, and $-1<\alpha<0$, $-1<\beta<0$. Now suppose the existence of functions φ_m and φ_{m-1} such that

$$\int_{-1}^1 g_m(x)\varphi_m(x)\,dx=\int_{-1}^1 g_{m-1}(x)\varphi_{m-1}(x)\,dx \tag{5.77}$$

for every f in L_1 in (5.75) and (5.76). Then we must have

$$\int_{-1}^1 \varphi_m(x)(x-t)^m k_{\alpha,\beta}(x-t)\,dx=\int_{-1}^1 \varphi_{m-1}(x)(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dx. \tag{5.78}$$

Now let c be chosen such that

$$\int_{-1}^{1}\{\varphi_m(x)-cW_{\alpha,\beta}(x)\}\,dx=0. \tag{5.79}$$

Then setting

$$\theta_m(x)=\varphi_m(x)-cW_{\alpha,\beta}(x), \qquad -1<x<1, \tag{5.80}$$

we have from (5.40) and (5.78)

$$\begin{aligned}\int_{-1}^{1}\varphi_{m-1}(x)(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dx-\int_{-1}^{1}\theta_m(x)(x-t)^m k_{\alpha,\beta}(x-t)\,dx\\ =c2^m P_m^{(\beta-m,\alpha-m)}(x), \qquad -1<x<1.\end{aligned} \tag{5.81}$$

Now since $\int_{-1}^{1}\theta_m(x)\,dx=0$, we have, integrating by parts,

$$\int_{-1}^{1}\theta_m(x)(x-t)^m k_{\alpha,\beta}(x-t)\,dx=\int_{-1}^{1}\Theta_m(x)(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dx \tag{5.82}$$

where

$$\Theta_m(x)=(\nu-m)\int_{-1}^{x}\theta_m(t)\,dt. \tag{5.83}$$

Hence we have

$$\begin{aligned}\int_{-1}^{1}\{\varphi_{m-1}(x)-\Theta_m(x)\}(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dx\\ =c2^m P_m^{(\beta-m,\alpha-m)}(x), \qquad -1<x<1.\end{aligned} \tag{5.84}$$

But from Theorem 5.57 we must have

$$cP_m^{(\beta-m,\alpha-m)}(x)=aP_{m-1}^{(\beta-m+1,\alpha-m+1)}(x)+bP_m^{(\beta-m+1,\alpha-m+1)}(x). \tag{5.85}$$

However, according to (5.74),

$$\begin{aligned}P_m^{(\beta^*,\alpha^*)}(x)=&\frac{(\nu+1-m)(\nu-m)}{\nu(\nu+1)}P_m^{(\beta^*+1,\alpha+1)}(x)\\ &+\frac{(\nu-m)(\beta-\alpha)}{(\nu-1)(\nu+1)}P_{m-1}^{(\beta^*+1,\alpha^*+1)}(x)-\frac{\alpha\beta}{\nu(\nu-1)}P_{m-2}^{(\beta^*+1,\alpha^*+1)}(x)\end{aligned} \tag{5.86}$$

where $\beta^*=\beta-m$, $\alpha^*=\alpha-m$, $\nu=\alpha+\beta+1$. Now for $-1<\alpha<0$ and $-1<\beta<0$, the leading coefficient of $P_k^{(\beta^*+1,\alpha^*+1)}$ does not vanish. So for $m\geq 2$ in (5.85) we must have $a=b=c=0$. Thus we have proved

THEOREM 5.87. *If the equation*

$$\int_{-1}^{1}f_m(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt=g(x), \qquad -1<x<1, \tag{5.88}$$

where $-1<\alpha<0$ $-1<\beta<0$, *m an integer* ≥ 2, *has a solution* f_m, *then*

$$\int_{-1}^{1}f_{m-1}(t)(x-t)^{m-1}k_{\alpha,\beta}(x-t)\,dt=g(x), \qquad -1<x<1, \tag{5.89}$$

has a solution f_{m-1} *if and only if*

$$\int_{-1}^{1}f_m(t)\,dt=0, \tag{5.90}$$

in which case

$$f_{m-1}(x)=(m-\nu)\int_{-1}^{x} f_m(t)\,dt. \tag{5.91}$$

6. An alternate solution to the first problem. Let us write

$$\sigma(x)\equiv\sigma(x;\alpha,\beta)=\int_{-1}^{1} W_{\beta,\alpha}(t)k_{\alpha,\beta}(x-t)\,dt, \tag{6.1}$$

$$-1<\alpha<0, \quad -1<\beta<0 \quad (\nu=\alpha+\beta+1\neq 0).$$

We have from (5.19)

$$\sigma(x)=1, \qquad -1<x<1. \tag{6.2}$$

For x outside $[-1,1]$ we can differentiate (6.1) and find that

$$\frac{d}{dx}\sigma(x)=\begin{cases}\dfrac{\nu}{\pi}2^{\nu}\sin\pi\alpha\dfrac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\nu+1)}|x+1|^{-\beta-1}|x-1|^{-\alpha-1}, & x>1,\\ -\dfrac{\nu}{\pi}2^{\nu}\sin\pi\beta\dfrac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\nu+1)}|x+1|^{-\beta-1}|x-1|^{-\alpha-1}, & x<-1.\end{cases} \tag{6.3}$$

We can formulate from (6.1)–(6.3) an alternate form of the solution to the first problem:

$$\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=g(x), \tag{6.4}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0, \quad \nu=\alpha+\beta+1\neq 0.$$

Let us define for $-1<\tau<1$

$$\varphi(\tau)=\int_{\tau}^{1} g(x)(x-\tau)^{\beta}(1-x)^{\alpha}dx. \tag{6.5}$$

We may interchange the order of integration, putting (6.4) in (6.5) to obtain

$$\varphi(\tau)=\int_{-1}^{1} f(t)\,dt\int_{\tau}^{1} k_{\alpha,\beta}(x-t)(x-\tau)^{\beta}(1-x)^{\alpha}dx. \tag{6.6}$$

If in the inner integral we put $x=(1+\tau)/2-(1-\tau)u/2$ we obtain

$$\int_{\tau}^{1} k_{\alpha,\beta}(x-t)(x-\tau)^{\beta}(1-x)^{\alpha}dx$$

$$=\int_{-1}^{1} k_{\alpha,\beta}\left(\frac{1+\tau}{2}-t-\frac{1-\tau}{2}u\right)\left(\frac{1-\tau}{2}\right)^{\nu}(1-u)^{\beta}(1+u)^{\alpha}du. \tag{6.7}$$

Then noting that $k_{\alpha,\beta}(ax)=a^{-\nu}k_{\alpha,\beta}(x)$ for $a>0$, we see from (6.1) that

$$\int_{\tau}^{1} k_{\alpha,\beta}(x-t)(x-\tau)^{\beta}(1-x)^{\alpha}dx=\sigma\left(\frac{1+\tau-2t}{1-\tau}\right), \qquad -1<\tau<1. \tag{6.8}$$

Hence

$$\varphi(\tau)=\int_{-1}^{1} f(t)\sigma\left(\frac{1+\tau-2t}{1-\tau}\right)dt. \tag{6.9}$$

Integrating by parts we obtain

$$\varphi(\tau)=\int_{-1}^{1} f(t)\,dt+\frac{2}{1-\tau}\int_{-1}^{1}\sigma'\left(\frac{1+\tau-2t}{1-\tau}\right)dt\int_{-1}^{t} f(x)\,dx. \tag{6.10}$$

From (6.2) and (6.3) we have

$$\sigma'\left(\frac{1+\tau-2t}{1-\tau}\right)=\begin{cases}0, & \tau<t<1,\\ \dfrac{(1-\tau)^{\nu+1}\nu\sin\pi\alpha\Gamma(\alpha+1)\Gamma(\beta+1)}{2\pi\Gamma(\nu+1)(1-t)^{\beta+1}(\tau-t)^{\alpha+1}}, & t<\tau<1.\end{cases} \tag{6.11}$$

Comparing (6.5) and (5.25) we see that

$$\int_{-1}^{1} f(t)\,dt=\varphi(-1). \tag{6.12}$$

Hence we have

$$\frac{\varphi(-1)-\varphi(\tau)}{(1-\tau)^{\nu}}=-\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\pi\Gamma(\nu)}\sin\pi\alpha\int_{-1}^{\tau}\frac{dt\int_{-1}^{t}f(x)\,dx}{(\tau-t)^{\alpha+1}(1-t)^{\beta+1}}. \tag{6.13}$$

This convolution equation (Abel's equation) is readily solved to obtain

$$\int_{-1}^{x} f(t)\,dt=\frac{(1-x)^{\beta+1}\Gamma(\nu)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\frac{d}{dx}\int_{-1}^{x}\frac{\varphi(-1)-\varphi(t)}{(1-t)^{\nu}(x-t)^{-\alpha}}\,dt \tag{6.14}$$

where $\varphi(x)=\int_x^1 g(t)(t-x)^{\beta}(1-t)^{\alpha}\,dt$.

In a similar fashion we find

$$\int_{x}^{1} f(t)\,dt=-\frac{(1+x)^{\alpha+1}\Gamma(\nu)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\frac{d}{dx}\int_{x}^{1}\frac{\psi(1)-\psi(t)}{(1+t)^{\nu}(t-x)^{-\beta}}\,dt \tag{6.15}$$

where $\psi(x)=\int_{-1}^{x} g(t)(x-t)^{\alpha}(1+t)^{\beta}\,dt$. (The connection between (6.14) and (6.15) is readily seen by changing the variables in (6.4) to obtain $\int_{-1}^{1} f(-t)k_{\beta,\alpha}\,dt=g(-x)$. Then one relation follows immediately from the other.)

By a change of variables we can write

$$\varphi(x)=(1-x)^{\nu}\int_0^1 u^{\beta}(1-u)^{\alpha}g[u+x(1-u)]\,du, \tag{6.16}$$

$$\psi(x)=(1+x)^{\nu}\int_0^1 u^{\alpha}(1-u)^{\beta}g(-u+x(1-u))\,du. \tag{6.17}$$

Then with the special pair $\{g(t)=1, f(t)=W_{\beta,\alpha}(t)\}$, we have from (6.14) and (6.15)

$$\begin{aligned}\int_{-1}^{x}(1-t)^{\beta}(1+t)^{\alpha}dt&=\frac{(1-x)^{\beta+1}}{\nu}\frac{d}{dx}\int_{-1}^{x}\left[\left(\frac{1-t}{2}\right)^{-\nu}-1\right](x-t)^{\alpha}dt\\ &=(1-x)^{\beta+1}2^{\nu}\int_{-1}^{x}(1-t)^{-\nu-1}(x-t)^{\alpha}dt, \qquad -1<x<1,\end{aligned} \tag{6.18}$$

$$\begin{aligned}\int_{x}^{1}(1-t)^{\beta}(1+t)^{\alpha}dt&=\frac{(1+x)^{\alpha+1}}{\nu}\frac{d}{dx}\int_{x}^{1}\left[\left(\frac{1+t}{2}\right)^{-\nu}-1\right](t-x)^{\beta}dt\\ &=(1+x)^{\alpha+1}2^{\nu}\int_{x}^{1}(1+t)^{-\nu-1}(t-x)^{\beta}dt, \qquad -1<x<1.\end{aligned} \tag{6.19}$$

The equations (6.18) and (6.19) are valid for $\alpha,\beta>-1$ (i.e., we may have $\nu\geq+1$), and may be verified by expressing the integrals in terms of hypergeometric functions and using transformation formulas for the functions.

If we set

$$f(x)=f_0(x)+c(1-x)^{\beta}(1+x)^{\alpha} \tag{6.20}$$

where

$$c=\frac{\Gamma(\nu+1)}{2^{\nu}\Gamma(\alpha+1)\Gamma(\beta+1)}\int_{-1}^{1}g(t)(1-t)^{\alpha}(1+t)^{\beta}dt, \tag{6.21}$$

we have

$$\int_{-1}^{1}f_0(t)\,dt=0. \tag{6.22}$$

Now assuming that we may differentiate inside the integrals in (6.14) and (6.15), using the elementary formulas

$$\frac{d}{dx}\int_{-1}^{x}P(t)(x-t)^{\alpha}dt=P(-1)(x+1)^{\alpha}+\int_{-1}^{x}P'(t)(x-t)^{\alpha}dt, \tag{6.23}$$

$$\frac{d}{dx}\int_{x}^{1}P(t)(t-x)^{\beta}dt=P(1)(1-x)^{\beta}+\int_{x}^{1}P'(t)(t-x)^{\beta}dt, \tag{6.24}$$

and (6.16)–(6.21), we obtain for $-1<x<1$:

(6.25)

$$\int_{-1}^{x}f_0(t)\,dt$$

$$=\frac{-(1-x)^{\beta+1}(1+x)^{\alpha+1}\int_0^1(1-t)^{\alpha}dt\int_0^1u^{\beta}(1-u)^{\alpha+1}g'\{2u-1+t(x+1)(1-u)\}\,du}{\nu(\nu+1)\left[\int_0^1(1-t)^{\alpha}dt\right]\left[\int_0^1u^{\beta}(1-u)^{\alpha+1}du\right]},$$

(6.26)

$$\int_{x}^{1}f_0(t)\,dt$$

$$=\frac{(1-x)^{\beta+1}(1+x)^{\alpha+1}\int_0^1(1-t)^{\beta}dt\int_0^1u^{\alpha}(1-u)^{\beta+1}g'\{1-2u-t(1-x)(1-u)\}\,du}{\nu(\nu+1)\left[\int_0^1(1-t)^{\beta}dt\right]\left[\int_0^1u^{\alpha}(1-u)^{\beta+1}du\right]}.$$

These formulas with (6.1) clearly show that if g agrees on $(-1,1)$ with a polynomial of degree n, then $f(t)/W_{\beta,\alpha}(t)$ agrees on $(-1,1)$ with a polynomial of degree n. Equivalent forms are

$$\int_{-1}^{x}f_0(t)\,dt=-\frac{(1-x)^{\beta+1}\Gamma(\nu)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\int_{-1}^{x}\frac{(x-t)^{\alpha}dt}{(1-t)^{\nu+1}}\int_{t}^{1}(1-s)^{\alpha+1}(s-t)^{\beta}g'(s)\,ds \tag{6.27}$$

$$(-1<x<1),$$

$$\int_{x}^{1}f_0(t)\,dt=\frac{(1+x)^{\alpha+1}\Gamma(\nu)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\int_{x}^{1}\frac{(t-x)^{\beta}dt}{(1+t)^{\nu+1}}\int_{-1}^{t}(1+s)^{\beta+1}(t-s)^{\alpha}g'(s)\,ds \tag{6.28}$$

$$(-1<x<1).$$

Now the hypergeometric function satisfies the integral equation (see for example, [4, p. 55])

$$F(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(\lambda)\Gamma(c-\lambda)}\int_0^1t^{\lambda-1}(1-t)^{c-\lambda-1}F(a,b;\lambda;zt)\,dt, \tag{6.29}$$

$$\mathrm{Re}\,c>\mathrm{Re}\,\lambda>0,\quad z\neq1,\quad |\arg(1-z)|<\pi.$$

Applying (6.29) to

$$P_n^{(\alpha,\beta)}(x)=\frac{\Gamma(n+1+\alpha)}{n!\Gamma(\alpha+1)}F\left(-n,n+\alpha+\beta+1;\alpha+1;\frac{1-x}{2}\right) \tag{6.30}$$
$$=(-1)^n P_n^{(\beta,\alpha)}(-x)$$

results in

$$\int_x^1(1-t)^\alpha(t-x)^\gamma P_n^{(\alpha,\beta)}(t)\,dt \tag{6.31}$$
$$=(1-x)^a\frac{\Gamma(\gamma+1)\Gamma(n+\alpha+1)}{\Gamma(n+a+1)}P_n^{(a,b)}(x)\qquad(-1<x<1)$$

where $a=\alpha+\gamma+1$, $b=\beta-\gamma-1$, and

$$\int_{-1}^x(1+t)^\beta(x-t)^\gamma P_n^{(\alpha,\beta)}(t)\,dt \tag{6.32}$$
$$=(1+x)^d\frac{\Gamma(\gamma+1)\Gamma(n+\beta+1)}{\Gamma(n+d+1)}P_n^{(c,d)}(x)\qquad(-1<x<1)$$

where $c=\alpha-\gamma-1$, $d=\beta+\gamma+1$. If we set

$$g(x)=\frac{(\nu)_n}{n!}P_n^{(\alpha,\beta)}(x),\qquad -1<x<1,\quad n=1,2,,\cdots, \tag{6.33}$$

we have from (5.44)

$$g'(x)=\frac{\Gamma(\nu+n+1)}{2\Gamma(\nu)n!}P_{n-1}^{(\alpha+1,\beta+1)}(x)\qquad(-1<x<1)\quad n=1,2,3,\cdots. \tag{6.34}$$

Then putting (6.34) in (6.27) we have from (6.31)

$$\frac{\Gamma(\nu)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\int_t^1(1-s)^{\alpha+1}(s-t)^\beta g'(x)\,ds \tag{6.35}$$
$$=\frac{(1-s)^{\nu+1}\Gamma(n+\alpha+1)}{2\Gamma(\alpha+1)n!}P_{n-1}^{(\nu+1,0)}(t),\qquad -1<t<1,\quad n=1,2,3,\cdots$$

Then applying (6.32) we obtain for g given by (6.33)

$$\int_{-1}^x f_0(t)\,dt=-\frac{1}{2n}(1-x)^{\beta+1}(1+x)^{\alpha+1}P_{n-1}^{(\beta+1,\alpha+1)}(x), \tag{6.36}$$
$$-1<x<1,\quad n=1,2,3,\cdots$$

It follows from (5.67) that in (6.36) we must have

$$f_0(t)=(1-t)^\beta(1+t)^\alpha P_n^{(\beta,\alpha)}(t),\qquad -1<t<1,\quad n=1,2,3,\cdots, \tag{6.37}$$

which is in agreement with (5.19), noting that in (6.20), $c=0$ for g given by (6.33).

As an example, we solve

$$\int_{-1}^1 f(t)k_{\alpha,\beta}(x-t)\,dt=\frac{(\nu)_n}{n!}\left(\frac{x-1}{2}\right)^n,\qquad -1<x<1,\quad n=1,2,3,\cdots. \tag{6.38}$$

Setting

$$f(t)=f_0(t)+\frac{(-1)^n\nu\Gamma(\alpha+n+1)}{(\nu+n)\Gamma(\alpha)n!}W_{\beta,\alpha}(t),\qquad -1<t<1, \tag{6.39}$$

we find from (6.27), (6.29), and (6.30),

$$\int_{-1}^{x} f_0(t)\,dt=\frac{(1-x)^{\beta+1}(1+x)^{\alpha+1}}{2(\nu+n)}P_{n-1}^{(-n-\alpha,\alpha+1)}(x), \tag{6.40}$$

$$-1<x<1, \quad n=1,2,3,\cdots.$$

Differentiating (6.40) and making use of the identities:

(6.41)

$$(1-x^2)\frac{d}{dx}P_{n-1}^{(-n-\alpha,\alpha+1)}(x)$$

$$=\{(n+1)x-n-1-2\alpha\}P_{n-1}^{(-n-\alpha,\alpha+1)}(x)-2nP_n^{(-n-\alpha,\alpha+1)}(x), \qquad n=1,2,3,\cdots,$$

$$(n+1)P_n^{(-n-\alpha,\alpha)}(x)=P_n^{(-n-\alpha,\alpha+1)}(x)-\alpha P_{n-1}^{(-n-\alpha,\alpha+1)}(x), \qquad n=1,2,3,\cdots, \tag{6.42}$$

(6.43)

$$(1+x)P_{n-1}^{(-n-\alpha,\alpha)}(x)=\frac{2(n+\alpha)}{n}P_{n-1}^{(-n-\alpha,\alpha)}(x)+2P_n^{(-n-\alpha,\alpha)}(x), \qquad n=1,2,3,\cdots,$$

where in (6.43) we have an exceptional case where a reduction of degree occurs, viz.,

(6.44)

$$P_{n-1}^{(-n-\alpha,\alpha)}(x)=(-1)^{n-1}P_{n-1}^{(\alpha,-n-\alpha)}(1)=(-1)^{n-1}\frac{\Gamma(\alpha+n)}{(n-1)!\Gamma(\alpha+1)}, \qquad n=1,2,3,\cdots,$$

we ultimately find

$$\int_{-1}^{1} W_{\beta,\alpha}(t)P_n^{(-n-\alpha,\alpha)}(t)k_{\alpha,\beta}(x-t)\,dt=\frac{(\nu)_n}{n!}\left(\frac{x-1}{2}\right)^n, \tag{6.45}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0, \quad \nu=\alpha+\beta+1, \quad n=0,1,2,\cdots.$$

Then from elementary transformations we have

$$\int_{-1}^{1} W_{\beta,\alpha}(t)P_n^{(\beta,-n-\beta)}(t)k_{\alpha,\beta}(x-t)\,dt=\frac{(\nu)_n}{n!}\left(\frac{1+x}{2}\right)^n, \tag{6.46}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0, \quad \nu=\alpha+\beta+1, \quad n=0,1,2,\cdots.$$

Actually, the manipulations leading to (6.45) were suggested by the formal condition that $n-1$ derivatives of the integral should vanish at $x=1$:

$$\int_{-1}^{1} W_{\beta,\alpha}(t)P_n(t)(1-t)^{-\nu-n+1+r}dt=0, \qquad r=0,1,2,\cdots,n-1,$$

which only makes sense for $r=n-1$, but suggests that $P_n=cP_n^{(-n-\alpha,\alpha)}$.

We note that $P_n^{(-n-\alpha,\alpha)}(x)$ has the fairly simple explicit representation

$$P_n^{(-n-\alpha,\alpha)}(x)=(-1)^n(\alpha+1)_n\sum_{k=0}^{n}\frac{(-1)^k((1+x)/2)^k}{(n-k)!(\alpha+1)_k}, \tag{6.47}$$

and similarly

$$P_n^{(\beta,-n-\beta)}(x)=(\beta+1)_n\sum_{k=0}^{n}\frac{(-1)^k((1-x)/2)^k}{(n-k)!(\beta+1)_k}. \tag{6.48}$$

Thus we find the solution of

$$\int_{-1}^{1}\left\{\sum_{0}^{n} b_k\left(\frac{1+t}{2}\right)^k\right\} W_{\beta,\alpha}(t)k_{\alpha,\beta}(x-t)\,dt \tag{6.49}$$

$$=\sum_{0}^{n}\binom{n}{k}\frac{(\nu)_k}{(\alpha+1)_k}a_k\left(\frac{1-x}{2}\right)^k, \qquad -1<x<1,$$

$$b_k=\frac{n!}{(n-k)!}\frac{(-1)^k}{(\alpha+1)_k}\sum_{j=0}^{n-k}\binom{n-k}{j}a_{k+j}. \tag{6.50}$$

7. Connection between solutions of equations with conjugate kernels. There is a fairly simple relation between f_1 and f_2 in

$$\int_{-1}^{1} f_1(t)k_{\alpha,\beta}(x-t)\,dt=g(x), \qquad -1<x<1, \tag{7.1}$$

$$\int_{-1}^{1} f_2(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=g(x), \qquad -1<x<1, \tag{7.2}$$

whenever both equations have solutions.

To see this we introduce the function analytic in the upper half plane defined by

$$F(z)=\int_{-1}^{1}\frac{f(t)\,dt}{\pi i(t-z)}, \qquad \operatorname{Im} z>0, \tag{7.3}$$

where we suppose initially that $f(t)$ is a real-valued function of L_1 which vanishes outside $[-1,1]$. We have on the boundary

$$F(x)\equiv F(x+i0)=f(x)+i\tilde{f}(x) \tag{7.4}$$

where

$$\tilde{f}(x)=\int_{-1}^{1}\frac{f(t)}{\pi(x-t)}dt. \tag{7.5}$$

We have the reciprocal relation

$$f(x)=\int_{-\infty}^{\infty}\frac{\tilde{f}(t)}{\pi(t-x)}dt \tag{7.6}$$

and, given only the projection of $\tilde{f}$ on $(-1,1)$ (Hilbert's problem), the "almost-reciprocal" relation

$$f(x)=(1-x^2)^{-1/2}\int_{-1}^{1}\frac{(1-t^2)^{1/2}\tilde{f}(t)}{\pi(t-x)}dt+\frac{c}{\pi}(1-x^2)^{-1/2}, \qquad -1<x<1. \tag{7.7}$$

Now consider the function analytic in the upper half plane

$$A(z)=W_{\beta,\alpha}(z)F(z)K_{\beta,\alpha}(z-\tau), \qquad -1<\tau<1, \tag{7.8}$$

where $W_{\beta,\alpha}$ and $K_{\beta,\alpha}$ are defined in (2.1), (2.2) and (3.13), (3.14) respectively. We see that $A(x)$ is real valued on the real axis outside $[-1,1]$ and $A(z)=O(z^2)$, $z\to\infty$. If we assume further that A is everywhere locally integrable on the real axis, then A belongs to L_1 on the real axis, so that

$$\int_{-\infty}^{\infty}A(t)\,dt=0, \tag{7.9}$$

$$\text{(7.10)}\qquad \operatorname{Im}\int_{-\infty}^{\infty} A(t)\,dt=\int_{-1}^{1}\operatorname{Im}A(t)\,dt=0,$$

$$\text{(7.11)}\qquad \int_{-1}^{1} W_{\beta,\alpha}(t)\tilde{f}(t)k_{\alpha,\beta}(\tau-t)\,dt=\int_{-1}^{1} W_{\beta,\alpha}(t)f(t)\tilde{k}_{\alpha,\beta}(\tau-t)\,dt,\qquad -1<\tau<1.$$

(Here we have used $k_{\beta,\alpha}(x)=k_{\alpha,\beta}(-x)$ and $\tilde{k}_{\beta,\alpha}(x)=-\tilde{k}_{\alpha,\beta}(-x)$.) Then setting

$$\text{(7.12)}\qquad f_2(t)=W_{\beta,\alpha}(t)f(t),\qquad -1<t<1,$$

$$\text{(7.13)}\qquad f_1(t)=W_{\beta,\alpha}(t)\tilde{f}(t),\qquad -1<t<1,$$

we have from (7.5) and (7.7) the relations

$$\text{(7.14)}\qquad f_1(x)=\int_{-1}^{1}\frac{W_{\beta,\alpha}(x)f_2(t)}{W_{\beta,\alpha}(t)\pi(x-t)}\,dt,\qquad -1<x<1,$$

$$\text{(7.15)}\qquad f_2(x)=\int_{-1}^{1}\frac{W_{\beta',\alpha'}(x)f_1(t)}{W_{\beta',\alpha'}(t)\pi(x-t)}\,dt+cW_{\beta',\alpha'}(x),\qquad -1<x<1,$$

where $\beta'=\beta-\frac{1}{2}$, $\alpha'=\alpha-\frac{1}{2}$ and c is an arbitrary constant whenever $W_{\beta',\alpha'}$ is integrable over $(-1,1)$.

8. Certain linear operator pairs. Here we assume we are given

$$\text{(8.1)}\qquad \int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=g(x),\qquad -1<x<1,$$

where $k_{\alpha,\beta}$ is a standard kernel of the first type with $-1<\alpha<0$ and $-1<\beta<0$, and we wish to establish certain pairs of linear operators A and B such that

$$\text{(8.2)}\qquad g^*(x)=(Ag)(x),\qquad -1<x<1,$$

$$\text{(8.3)}\qquad f^*(x)=\begin{cases}(Bf)(x), & -1<x<1,\\ 0, & |x|>1,\end{cases}$$

$$\text{(8.4)}\qquad \int_{-1}^{1} f^*(t)k_{\alpha,\beta}(x-t)\,dt=g^*(x),\qquad -1<x<1.$$

8.1 Integral operators: $g^*(x)=\int_0^x g(t)\,dt$. Here we define

$$\text{(8.1.1)}\qquad f_0(t)=\begin{cases}f(t)-aW_{\beta,\alpha}(t), & -1<t<1,\\ 0, & |t|>1,\end{cases}$$

where the constant a is chosen so that $\int_{-\infty}^{\infty} f_0\,dt=0$; i.e., in accordance with (5.25),

$$\text{(8.1.2)}\qquad \int_{-1}^{1}(g(t)-a)W_{\alpha,\beta}(t)\,dt=0,$$

we have

$$\text{(8.1.3)}\qquad \int_{-\infty}^{\infty} k_{\alpha,\beta}(t)f_0(x-t)\,dt=g(x)-a,\qquad -1<x<1,$$

and hence,

$$\text{(8.1.4)}\qquad \int_{-\infty}^{\infty} k_{\alpha,\beta}(t)\{F(x-t)-F(-t)\}\,dt=\int_0^x g(t)\,dt-ax,\qquad -1<x<1,$$

where

$$F(x)=\int_{-\infty}^{x} f_0(t)\,dt, \tag{8.1.5}$$

and according to (8.1.1), $F(x)=0$ for $|x|>1$. Thus, rewriting (8.1.1), we have

$$\int_{-1}^{1} F(t)k_{\alpha,\beta}(x-t)\,dt=\int_{0}^{x} g(t)\,dt-ax+b \tag{8.1.6}$$

where the constant b is determined, once again in accordance with (5.25), by

$$\int_{-1}^{1} F(t)\,dt=\int_{-1}^{1} W_{\alpha,\beta}(x)\left\{\int_{0}^{x} g(t)\,dt-ax+b\right\}dx. \tag{8.1.7}$$

Now

$$\int_{-1}^{1} F(t)\,dt=-\int_{-1}^{1} tf_0(t)\,dt$$

and

$$\int_{-1}^{1}\{g(t)-a\}W_{\alpha,\beta}(t)P_1^{(\alpha,\beta)}(t)\,dt=\nu\int_{-1}^{1} P_1^{(\beta,\alpha)}(t)f_0(t)\,dt,$$

or, since

$$\int_{-1}^{1} f_0(t)\,dt=0=\int_{-1}^{1}\{g(t)-a\}W_{\alpha,\beta}(t)\,dt,$$

$$\nu\int_{-1}^{1} tf_0(t)\,dt=\int_{-1}^{1} t\{g(t)-a\}W_{\alpha,\beta}(t)\,dt \tag{8.1.8}$$

(cf. (5.6)). Thus the constant b is determined from the projection of g by

$$\int_{-1}^{1} W_{\alpha,\beta}(x)\left\{\int_{0}^{x} g(t)\,dt+\frac{x}{\nu}g(x)-a\left(1+\frac{1}{\nu}\right)x+b\right\}dx=0 \tag{8.1.9}$$

together with (8.1.2). Finally, we have from (8.1.6) and (5.30)–(5.33)

$$\int_{-1}^{1} f^*(t)k_{\alpha,\beta}(x-t)\,dt=\int_{0}^{x} g(t)\,dt, \qquad -1<x<1, \tag{8.1.10}$$

where

(8.1.11)

$$f^*(x)=\int_{-1}^{x}\{f(t)-aW_{\beta,\alpha}(t)\}\,dt+W_{\beta,\alpha}(x)\left\{\frac{a}{\nu}(\alpha-\beta+x)-b\right\}, \qquad -1<x<1,$$

and $\nu=\alpha+\beta+1$, and the constants a and b are determined by (8.1.2) and (8.1.9).

8.2. Differential operator: $g^*(x)=\frac{d}{dx}g(x)$. If we assume the equation

$$\int_{-1}^{1} f^*(t)k_{\alpha,\beta}(x-t)\,dt=\frac{d}{dx}g(x), \tag{8.2.1}$$

$$-1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0$$

has a solution f^*, then as we have seen in the previous section, the equation

$$\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=g(x), \qquad -1<x<1, \quad -1<\alpha<0, \quad -1<\beta<0, \tag{8.2.2}$$

always has a solution f, the relation being

$$f^*(x)=aW_{\beta,\alpha}(x)+\frac{d}{dx}\{f(x)-W_{\beta,\alpha}(x)P_1(x)\}, \qquad -1<x<1, \tag{8.2.3}$$

where

$$P_1(x)=\frac{a}{\nu}(\alpha-\beta+x)-b \quad \text{with}\ \nu=\alpha+\beta+1 \tag{8.2.4}$$

and a and b are determined by

$$\int_{-1}^{1}\{g'(t)-a\}W_{\alpha,\beta}(t)\,dt=0, \tag{8.2.5}$$

$$\int_{-1}^{1}\left\{g(t)+\frac{t}{\nu}g'(t)-a\left(1+\frac{1}{\nu}\right)t+b\right\}W_{\alpha,\beta}(t)\,dt=0. \tag{8.2.6}$$

That is, if (8.2.2) has a solution f, then (8.2.1) has a solution f^* in L_1 if and only if there exists a polynomial P_1 of first degree such that $F(t)$ given by

$$F(t)=\xi(t)\{f(t)-W_{\beta,\alpha}(t)P_1(t)\} \tag{8.2.7}$$

where $\xi(t)$ is the characteristic function of the interval $(-1,1)$, is the integral of a function of L_1.

This condition then gives an alternate means of determining the constants a and b in (8.2.4) directly from f.

8.3 Multiplication by x: $f^*(x)=xf(x)$. We have the given relation (8.1) between f and g and seek g^* in

$$\int_{-1}^{1}tf(t)k_{\alpha,\beta}(x-t)\,dt=g^*(x), \qquad -1<x<1. \tag{8.3.1}$$

Then

$$\int_{-1}^{1}f(t)(x-t)k_{\alpha,\beta}(x-t)\,dt=xg(x)-g^*(x), \qquad -1<x<1. \tag{8.3.2}$$

Differentiating (8.3.2) with respect to x, using $\frac{d}{dx}xk_{\alpha,\beta}=-(\nu-1)k_{\alpha,\beta}$, where $\nu-1=\alpha+\beta$, we have

$$\frac{d}{dx}\{xg(x)-g^*(x)\}=-(\nu-1)g(x), \qquad -1<x<1. \tag{8.3.3}$$

Thus

$$g^*(x)=xg(x)+(\nu-1)\int_0^x g(t)\,dt+c, \qquad -1<x<1. \tag{8.3.4}$$

The constant c is determined by the relations (cf. (5.4)–(5.7) and (5.25))

$$\int_{-1}^{1}g^*(x)W_{\alpha,\beta}(x)\,dx=\int_{-1}^{1}tf(t)\,dt \tag{8.3.5}$$

and

$$\int_{-1}^{1}tf(t)\,dt=\frac{1}{\nu}\int_{-1}^{1}(t+\alpha-\beta)W_{\alpha,\beta}(t)g(t)\,dt. \tag{8.3.6}$$

8.4. Multiplication by x: $g^*(x)=xg(x)$. Once again we are given the relation (8.1), and now seek f^* in

$$\int_{-1}^{1}f^*(t)k_{\alpha,\beta}(x-t)\,dt=xg(x), \qquad -1<x<1. \tag{8.4.1}$$

We have from (8.3.1) and (8.3.4)

$$\int_{-1}^{1}tf(t)k_{\alpha,\beta}(x-t)\,dt=xg(x)+(\nu-1)\int_0^x g(t)\,dt+c, \qquad -1<x<1. \tag{8.4.2}$$

Also, from (8.1.11), we have

$$\int_{-1}^{1} f_1^*(t)k_{\alpha,\beta}(x-t)\,dt=\int_0^x g(t)\,dt, \qquad -1<x<1, \tag{8.4.3}$$

where

$$f_1^*(x)=\int_{-1}^{x}\{f(t)-aW_{\beta,\alpha}(t)\}\,dt+W_{\beta,\alpha}(x)\left\{\frac{a}{\nu}(\alpha-\beta+x)-b\right\}, \qquad -1<x<1. \tag{8.4.4}$$

Thus f^* in (8.4.1) is given by

$$f^*(x)=xf(x)-(\nu-1)\int_{-1}^{x}\{f(t)-aW_{\beta,\alpha}(t)\}\,dt-W_{\beta,\alpha}(x)\left\{\frac{\nu-1}{\nu}ax+b'\right\}, \qquad -1<x<1. \tag{8.4.5}$$

where the constants a and b' are determined by

$$\int_{-1}^{1}\{f(t)-aW_{\beta,\alpha}(t)\}\,dt=0=\int_{-1}^{1}\{g(t)-a\}W_{\alpha,\beta}(t)\,dt, \tag{8.4.6}$$

$$\int_{-1}^{1}f^*(t)\,dt=\int_{-1}^{1}tg(t)W_{\alpha,\beta}(t)\,dt=\int_{-1}^{1}(\nu t+\beta-\alpha)f(t)\,dt. \tag{8.4.7}$$

Integrating by parts one of the terms in f^*, we have

$$\int_{-1}^{1}f^*(t)\,dt=\int_{-1}^{1}\nu tf(t)\,dt-\int_{-1}^{1}W_{\beta,\alpha}(t)\left\{\frac{\nu^2-1}{\nu}at+b'\right\}dt. \tag{8.4.8}$$

Since $P_1^{\beta,\alpha}(t)=(\nu+1)t/2+(\beta-\alpha)/2$ we have

$$\int_{-1}^{1}(\nu+1)tW_{\beta,\alpha}(t)\,dt=(\alpha-\beta)\int_{-1}^{1}W_{\beta,\alpha}(t)\,dt. \tag{8.4.9}$$

So

$$\int_{-1}^{1}W_{\beta,\alpha}(t)\frac{\nu^2-1}{\nu}at\,dt=\frac{\nu-1}{\nu}(\alpha-\beta)a\int_{-1}^{1}W_{\beta,\alpha}(t)\,dt=\frac{\alpha^2-\beta^2}{\nu}\int_{-1}^{1}f(t)\,dt. \tag{8.4.10}$$

Putting (8.4.10) and (8.4.8) in (8.4.7) we have

$$\int_{-1}^{1}f(t)\,dt=\frac{\alpha^2-\beta^2}{\nu}\int_{-1}^{1}f(t)\,dt+b'\int_{-1}^{1}W_{\beta,\alpha}(t)\,dt \tag{8.4.11}$$

or

$$b'\int_{-1}^{1}W_{\beta,\alpha}(t)\,dt=\frac{\alpha-\beta}{\nu}\int_{-1}^{1}f(t)\,dt. \tag{8.4.12}$$

Therefore

$$b'=\frac{\alpha-\beta}{\nu}a \tag{8.4.13}$$

and

$$f^*(x)=xf(x)-(\nu-1)\int_{-1}^{x}\{f(t)-aW_{\beta,\alpha}(t)\}\,dt -\frac{a}{\nu}\{(\alpha+\beta)x+\alpha-\beta\}W_{\beta,\alpha}(x), \qquad -1<x<1, \tag{8.4.14}$$

where the constant a is determined by (8.4.6).

8.5. $f^*(x)=\frac{d}{dx}\{(1-x^2)f(x)\}+nxf(x)$. We have from §8.3,

$$\int_{-1}^{1} tf(t)k_{\alpha,\beta}(x-t)\,dt=xg(x)+(\nu-1)\int_0^x g(t)\,dt+c_1, \qquad -1<x<1. \tag{8.5.1}$$

Iteration of this formula gives

$$\begin{aligned}&\int_{-1}^{1} t^2f(t)k_{\alpha,\beta}(x-t)\,dt\\ &=x^2g(x)+(\nu-1)x\int_0^x g(t)\,dt+c_1x\\ &\quad+(\nu-1)\int_0^x\left\{tg(t)+(\nu-1)\int_0^t g(u)\,du+c_1\right\}dt+c_2, \qquad -1<x<1.\end{aligned} \tag{8.5.2}$$

Hence

$$\begin{aligned}&\int_{-1}^{1}(1-t^2)f(t)k_{\alpha,\beta}(x-t)\,dt\\ &=(1-x^2)g(x)-(\nu-1)x\int_0^x g(t)\,dt-c_1x\\ &\quad-(\nu-1)\int_0^x\left\{tg(t)+(\nu-1)\int_0^t g(u)\,du+c\right\}dt-c_2, \qquad -1<x<1.\end{aligned} \tag{8.5.3}$$

Now if we differentiate the right-hand side we have to replace $(1-t^2)f(t)$ by $\frac{d}{dt}\{(1-t^2)f(t)+P_1(t)W_{\beta,\alpha}(t)\}$, and it is clear that $P_1\equiv 0$, otherwise the derivative of the quantity in braces could not belong to $L_1(-1,1)$, given that f belongs to $L_1(-1,1)$. Thus we have

(8.5.4)

$$\begin{aligned}&\int_{-1}^{1} dtk_{\alpha,\beta}(x-t)\,\frac{d}{dt}\{(1-t^2)f(t)\}\\ &\quad=\frac{d}{dx}\{(1-x^2)g(x)\}-2(\nu-1)xg(x)-\nu(\nu-1)\int_0^x g(t)\,dt-\nu c_1, \qquad -1<x<1.\end{aligned}$$

Then if we multiply (8.5.1) by ν and add it to (8.5.4) we have

$$\int_{-1}^{1} f^*(t)k_{\alpha,\beta}(x-t)\,dt=g^*(x), \tag{8.5.5}$$

where

$$f^*(x)=\frac{d}{dx}\{(1-x^2)f(x)\}+\nu xf(x)=(1-x^2)f'(x)+(\nu-2)xf(x), \tag{8.5.6}$$

$$g^*(x)=\frac{d}{dx}\{(1-x^2)g(x)\}+(2-\nu)xg(x)=(1-x^2)g'(x)-\nu xg(x). \tag{8.5.7}$$

These formulas are valid provided $(1-x^2)f'(x)$ belongs to $L_1(-1,1)$. We note the equivalent forms:

$$f^*(x)=(1-x^2)^{\nu/2}\frac{d}{dx}(1-x^2)^{1-\nu/2}f(x), \tag{8.5.8}$$

$$g^*(x)=(1-x^2)^{1-\nu/2}\frac{d}{dx}(1-x^2)^{\nu/2}g(x). \tag{8.5.9}$$

More generally, if we set

$$\varphi^*(x)=(1-x)^{\mu}(1+x)^{\lambda}\frac{d}{dx}(1-x)^{1-\mu}(1+x)^{1-\lambda}f(x) \tag{8.5.10}$$
$$=(1-x^2)f'(x)-(1-\lambda-\mu)xf(x)+(\mu-\lambda)f(x)$$

and

$$\gamma^*(x)=(1-x)^{1-\lambda}(1+x)^{1-\mu}\frac{d}{dx}(1-x)^{\lambda}(1+x)^{\mu}g(x) \tag{8.5.11}$$
$$=(1-x^2)g'(x)-(\lambda+\mu)xg(x)+(\mu-\lambda)g(x)$$

where

$$\lambda+\mu=\nu, \tag{8.5.12}$$

we have

$$\int_{-1}^{1}\varphi^*(t)k_{\alpha,\beta}(x-t)\,dt=\gamma^*(x), \qquad -1<x<1 \tag{8.5.13}$$

whenever (8.1) holds and $(1-x^2)f'(x)$ belongs to $L_1(-1,1)$.

In particular, if $f=(1-x)^{\mu-1}(1+x)^{\lambda-1}$, $\mu>0$, $\lambda>0$, and $\mu+\lambda=\nu<1$, then $\varphi^*(x)\equiv 0$ and hence $\gamma^*\equiv 0$; i.e., $g=c(1-x)^{-\lambda}(1+x)^{-\mu}$, a result obtained previously [cf. (4.18)].

We note that the results of this section depend on α and β only through their sum; i.e., the results should hold for a kernel of the second type, and indeed they do, since the only use made of the kernel being of the first type was in finding solutions when first degree polynomials appeared with g, and these cancelled out in the operation here.

8.6. Weighted Hilbert transform. Here we are given the original relation (8.1) and set

$$g^*(x)=\frac{1}{W_{\alpha,\beta}(x)}\int_{-1}^{1}\frac{g(t)W_{\alpha,\beta}(t)}{\pi(x-t)}\,dt, \qquad -1<x<1. \tag{8.6.1}$$

From (3.26) and (3.4) we have

$$\int_{-1}^{1}f(t)\tilde{k}_{\alpha,\beta}(x-t)\,dt=g^*(x), \qquad -1<x<1. \tag{8.6.2}$$

Now we seek f^* in

$$\int_{-1}^{1}f^*(t)k_{\alpha,\beta}(x-t)\,dt=g^*(x), \qquad -1<x<1. \tag{8.6.3}$$

But if (8.6.3) has a solution f^*, we have seen in §7 that it is given by

$$f^*(x)=W_{\beta,\alpha}(x)\int_{-1}^{1}\frac{f(t)\,dt}{W_{\beta,\alpha}(t)\pi(x-t)}, \qquad -1<x<1. \tag{8.6.4}$$

9. Integral representation of certain analytic functions. Suppose $G(z)$ is analytic in the upper half plane and satisfies

$$\lim_{z\to\infty}G(z)=0 \qquad (\mathrm{Im}\,z\geq 0), \tag{9.1}$$

and on the real axis

$$G(x)=\begin{cases}r(x)e^{i\pi(\alpha+1/2)}, & x>1,\\ r(x)e^{-i\pi(\beta+1/2)}, & x<-1,\end{cases} \tag{9.2}$$

where $r(x)$ is a real-valued function, and $-1<\alpha<0$, $-1<\beta<0$, $0<\alpha+\beta+1<1$. Suppose further that for any finite T

$$\int_{-T}^{T}|G(x)|^p dx<\infty, \qquad 1<p<\frac{1}{\nu} \quad (\nu=\alpha+\beta+1). \tag{9.3}$$

Then, as argued in §3, G has the representation

$$G(z)=(1-z)^{-\alpha}(1+z)^{-\beta}\int_{-1}^{1}\frac{(1-t)^{\alpha}(1+t)^{\beta}g(t)}{\pi i(t-z)}dt \tag{9.4}$$

where we take the branch of $(1-z)^{\alpha}(1+z)^{\beta}$ that is real and positive in $(-1,1)$, and z is restricted to the upper half plane, and

$$G(x)\equiv g(x)+i\tilde{g}(x) \tag{9.5}$$

where g and $\tilde{g}$ are real-valued functions. It was shown in §3 that (9.3) implies

$$\int_{-1}^{1}\{(1-t)^{\alpha}(1+t)^{\beta}|g(t)|^p dt<\infty \quad \text{for } 1\le p<(\nu+\rho)^{-1} \tag{9.6}$$

where $\rho=\max(-\alpha,-\beta)$. Since $(1-t)^{\alpha}(1+t)^{\beta}g(t)$ is integrable over $(-1,1)$ we see from the representation (9.4) that (9.1) can be replaced by

$$G(z)=O(|z|^{-\nu}) \quad \text{as } z\to\infty. \tag{9.7}$$

In fact,

$$G(z)=e^{i\pi(\alpha+1/2)}z^{-\nu}(1-z^{-1})^{-\alpha}(1+z^{-1})^{-\beta}\int_{-1}^{1}\frac{(1-t)^{\alpha}(1+t)^{\beta}g(t)\,dt}{1-tz^{-1}}, \tag{9.8}$$

so that we have

$$G(z)=e^{i\pi(\alpha+1/2)}z^{-\nu}\sum_{n=0}^{\infty}a_n z^{-n} \tag{9.9}$$

where the a_n are real and the sum converges for $|z|>1$.

Now we seek the representation

$$G(z)=\int_{-\infty}^{\infty}f(t)K_{\alpha,\beta}(z-t)\,dt \tag{9.10}$$

where f is a *real-valued*[3] function of L_1, and we recall that

$$K_{\alpha,\beta}(z)=\frac{1}{\pi}e^{i\pi(\alpha+1/2)}z^{-\nu}, \qquad 0\le\arg z\le\pi \quad (\nu=\alpha+\beta+1). \tag{9.11}$$

Now if (9.10) has a solution f which is real valued, then f must vanish outside $[-1,1]$, for we have

$$e^{i\pi\beta}G(z)=\frac{1}{\pi}e^{i\pi(\nu-1/2)}\int_{-\infty}^{\infty}f(t)(z-t)^{-\nu}dt, \tag{9.12}$$

and for real x

$$e^{i\pi\beta}G(x)=\frac{1}{\pi}e^{i\pi(\nu-1/2)}\int_{-\infty}^{x}(x-t)^{-\nu}f(t)\,dt-\frac{i}{\pi}\int_{x}^{\infty}(t-x)^{-\nu}f(t)\,dt, \tag{9.13}$$

$$\operatorname{Re}e^{i\pi\beta}G(x)=\frac{1}{\pi}\sin\pi\nu\int_{-\infty}^{x}(x-t)^{-\nu}f(t)\,dt. \tag{9.14}$$

[3]Without this requirement the solution to (9.10) is not unique since the Fourier transform of $K_{\alpha,\beta}$ vanishes on a half line.

Similarly

$$e^{-i\pi\alpha}G(z)=\frac{i}{\pi}\int_{-\infty}^{\infty}f(t)(z-t)^{-\nu}dt. \tag{9.15}$$

and for real x

$$e^{-i\pi\alpha}G(x)=\frac{ie^{-i\pi\nu}}{\pi}\int_{-x}^{\infty}(t-x)^{-\nu}f(t)\,dt+\frac{i}{\pi}\int_{-\infty}^{x}f(t)(x-t)^{-\nu}dt, \tag{9.16}$$

$$\operatorname{Re}e^{-i\pi\alpha}G(x)=\frac{1}{\pi}\sin\pi\nu\int_{x}^{\infty}(t-x)^{-\nu}f(t)\,dt. \tag{9.17}$$

Now by supposition (9.2), $\operatorname{Re}e^{i\pi\beta}G(x)=0$ for $x<-1$, and $\operatorname{Re}e^{-i\pi\alpha}G(x)=0$ for $x>-1$. It follows from (9.14) and (9.17) that $f(t)=0$, $|t|>1$. Thus if (9.10) has a (real-valued) solution f, it is determined by

$$\int_{-1}^{1}f(t)k_{\alpha,\beta}(x-t)\,dt=g(x),\qquad -1<x<1, \tag{9.18}$$

and from (9.13) and (9.17), using the fact that $f(t)=0$, $|t|>1$, we have

$$\int_{-1}^{x}f(t)\,dt=\int_{-1}^{x}(x-t)^{\nu-1}g_1(t)\,dt, \tag{9.19}$$

where $g_1(t)=\operatorname{Re}e^{i\pi\beta}G(t)=\cos\pi\beta g(t)-\sin\pi\beta\tilde{g}(t)$, and

$$\int_{x}^{1}f(t)\,dt=\int_{x}^{1}(t-x)^{\nu-1}g_2(t)\,dt, \tag{9.20}$$

where $g_2(t)=\operatorname{Re}e^{-i\pi\alpha}G(t)=\cos\pi\alpha g(t)+\sin\pi\alpha\tilde{g}(t)$.

Now if g_1 or g_2 is a function of bounded variation, we may write, respectively,

$$f(x)=\int_{-1-0}^{x}(x-t)^{\nu-1}dg_1(t),\qquad -1<x<1, \tag{9.21}$$

$$f(x)=-\int_{x}^{1+0}(t-x)^{\nu-1}dg_2(t),\qquad -1<x<1. \tag{9.22}$$

Then f will belong to L_p for $1\le p<(1-\nu)^{-1}$. It is not necessary, of course, for g_1 or g_2 to be functions of bounded variation in order for (9.18) to have a solution f in L_1. We note that the convolution kernels in (9.19) and (9.20) can be written as the convolution of two kernels of the same form with exponents whose sum is $\nu-2$; i.e., for $0<\varepsilon<\nu$ we have

$$f(x)=\frac{d}{dx}\frac{1}{\Gamma(\varepsilon)}\int_{-1}^{x}(x-t)^{\varepsilon-1}g_3(t)\,dt, \tag{9.23}$$

$$f(x)=-\frac{d}{dx}\frac{1}{\Gamma(\varepsilon)}\int_{x}^{1}(t-x)^{\varepsilon-1}g_4(t)\,dt \tag{9.24}$$

where g_3 and g_4 (depending on ε) are given by

$$g_3(x)=\frac{\Gamma(\nu)}{\Gamma(\nu-\varepsilon)}\int_{-1}^{x}(x-t)^{\nu-\varepsilon-1}g_1(t)\,dt, \tag{9.25}$$

$$g_4(x)=\frac{\Gamma(\nu)}{\Gamma(\nu-\varepsilon)}\int_{x}^{1}(t-x)^{\nu-\varepsilon-1}g_2(t)\,dt. \tag{9.26}$$

Hence, if for some ε $(0<\varepsilon<\nu)$, g_3 or g_4 is a function of bounded variation, f will belong to L_p for $1\le p<(1-\varepsilon)^{-1}$. The condition, however, that g_1 or g_2 be a function of bounded variation is simply interpreted in terms of the mapping properties of $G(z)$;

i.e., if G maps the upper half plane one–one onto the interior of a simply connected region with a rectifiable boundary, then $G(x)$, and hence $g_1(x)$ and $g_2(x)$, are functions of bounded variation. Thus we can state the following theorem after one more observation concerning the positivity of f.

Suppose $-1<a<1$ and $g_1(x)$ is nondecreasing in $(-\infty,a)$ and $g_2(x)$ is nonincreasing in (a,∞) (Recall that $g_1(x)\equiv 0$ for $x<-1$, and $g_2(x)\equiv 0$ for $x>1$.). Then we see from (9.21) that $f(x)>0$ for $-1<x<a$, and from (9.22) that $f(x)>0$ for $a<x<1$.

THEOREM 9.27. *Suppose* $-1<\alpha<0$, $-1<\beta<0$, *and* $0<\nu<1$, *where* $\nu=\alpha+\beta+1$, *and* $G(z)$ *is an analytic function which maps the upper half plane one–one onto the interior of the region depicted in Fig.* 1 *where O is the origin and the image of* $z=\infty$; *the straight line segment OA is the image of* $(-\infty,-1)$ *and makes an angle* $(\beta+\frac{1}{2})\pi$, *measured positively in the clockwise direction, with the real axis OR; the wavy line AB is a simple rectifiable curve and is the image of* $(-1,1)$; *the straight line segment BO is the image of* $(1,\infty)$ *and makes an angle* $(\alpha+\frac{1}{2})\pi$, *measured positively in the counter-clockwise direction, with the real axis OR, and hence makes an angle* $\nu\pi$ *with OA. (The arrows on the boundary correspond to increasing x in* $G(x)$, $-\infty<x<\infty$.)

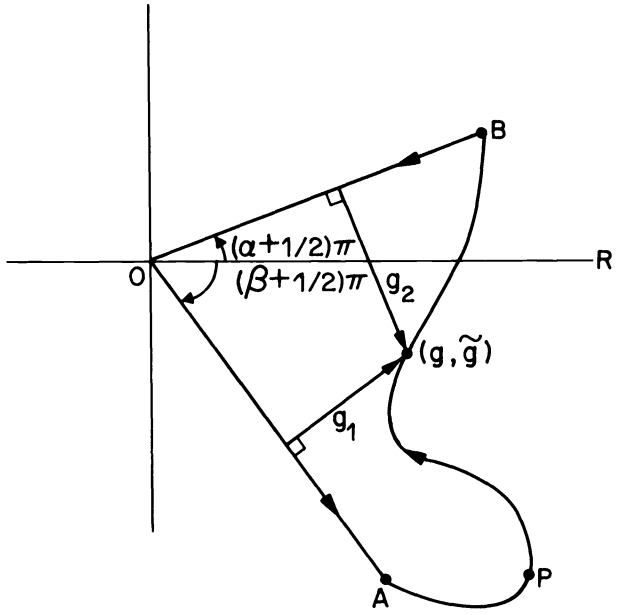


FIG. 1.

Then $G(z)$ *has the representation*

$$\text{i)}\qquad G(z)=\frac{i}{\pi}e^{i\pi\alpha}\int_{-1}^{1}f(t)(z-t)^{-\nu}dt,\qquad \operatorname{Im}z\geq 0,\quad 0\leq\arg z\leq\pi,$$

where f is a real-valued function given by (9.21) *or equivalently by* (9.22), *and*

$$\text{ii)}\qquad \int_{-1}^{1}|f(t)|^{p}dt<\infty\quad\text{for } 1\leq p<(1-\nu)^{-1}.$$

Furthermore, from (5.25) *we have*

$$\text{iii)}\qquad \int_{-1}^{1}(1-t)^{\alpha}(1+t)^{\beta}P_n^{(\alpha,\beta)}(t)g(t)\,dt=\frac{(\nu)_n}{n!}\int_{-1}^{1}P_n^{(\beta,\alpha)}(t)f(t)\,dt,$$

$$n=0,1,2,\cdots,$$

where $g(x)=\operatorname{Re}G(x)$.

Now suppose further that there is a point P on AB (cf. Fig. 1) which is the image of, say, $x=a$ $(-1<a<1)$ such that $g_1(x)\equiv \operatorname{Re} e^{i\pi\beta}G(x)$ is a nondecreasing function of x for $x<a$, and $g_2(x)\equiv \operatorname{Re} e^{i\pi\alpha}G(x)$ is a nonincreasing function of x for $x>a$. Then

iv) $$f(x)>0, \qquad -1<x<1.$$

Example 1. The function $G(z)=(\sqrt{1-z^2}+iz)^\nu$, where $0<\nu<1$ and we take the branch of $\sqrt{1-z^2}$ analytic in the upper half plane and positive in $(-1,1)$, maps the upper half plane one–one onto that sector of the unit disk with vertex angle $\nu\pi$ which is bisected by the positive real axis; i.e., in Fig. 1 $\alpha=\beta=\frac{1}{2}(\nu-1)$, and AB is the arc of a circle of radius 1. We have

$$G(x)=e^{i\nu\theta(x)}, \qquad -1\le x\le 1, \tag{9.28}$$

where $\theta(x)=\arcsin x$ $(-\frac{\pi}{2}<\theta<\frac{\pi}{2})$,

$$G(x)=\begin{cases} e^{i\nu\pi/2}\left(x-\sqrt{x^2-1}\right), & x>1, \\ e^{-i\nu\pi/2}\left(-x-\sqrt{x^2-1}\right), & x<-1, \end{cases} \tag{9.29}$$

where $\sqrt{x^2-1}>0$. Thus G has the representation

$$G(z)=e^{i\pi\nu/2}\int_{-1}^{1} f_\nu(t)(z-t)^{-\nu}dt, \qquad \operatorname{Im} z\ge 0, \quad 0\le \arg z\le\pi, \tag{9.30}$$

where according to (9.21), f_ν is given by

$$f_\nu(x)=\int_{-1}^{x}(x-t)^{\nu-1}\frac{d}{dt}\left\{\sin(\nu\theta(t))+\frac{\nu\pi}{2}\right\}dt \tag{9.31}$$

or

$$f_\nu(-\cos\varphi)=\nu\int_0^\varphi \frac{\cos\nu x\,dx}{(\cos x-\cos\varphi)^{1-\nu}}, \qquad 0\le\varphi\le\pi. \tag{9.32}$$

We have the representation of the Gegenbauer polynomials (cf. [4, p. 224]):

$$C_n^\nu(\cos\varphi)=\frac{2^{1-\nu}\Gamma(n+2\nu)}{n!\{\Gamma(\nu)\}^2}(\sin\varphi)^{1-2\nu}\int_0^\varphi \frac{\cos\{(n+\nu)x\}\,dx}{(\cos x-\cos\varphi)^{1-\nu}}, \qquad \nu>0. \tag{9.33}$$

It follows that

$$f_\nu(x)=\frac{\nu\{\Gamma(\nu)\}^2}{2^{1-\nu}\Gamma(2\nu)}(1-x^2)^{\nu-1/2}, \qquad -1<x<1. \tag{9.34}$$

Example 2. Let $G(z)$ be the analytic function which maps the upper half plane one–one onto the interior of the triangle depicted in Fig. 1 when AB becomes a vertical straight line segment at a unit distance from the origin. Then $G(z)$ has the representation

$$G(z)=\int_{-1}^{1} f(t)K_{\alpha,\beta}(z-t)\,dt \tag{9.35}$$

where f is the solution of

$$\int_{-1}^{1} f(t)k_{\alpha,\beta}(x-t)\,dt=1, \qquad -1<x<1. \tag{9.36}$$

Therefore, as we have seen,

$$f(t)=(1-t)^{\beta}(1+t)^{\alpha}, \qquad -1<t<1. \tag{9.37}$$

Examples 1 and 2 are special cases of some important formulas we give below. From (9.33) we can deduce that

$$\frac{1}{\pi}\int_{-1}^{1}(1-t^2)^{\nu-1/2}C_n^{\nu}(t)e^{i\nu\pi/2}(z-t)^{-\nu}dt=\frac{2^{1-\nu}\Gamma(n+2\nu)(-i)^n}{n!(n+\nu)\{\Gamma(\nu)\}^2}\left\{\sqrt{1-z^2}+iz\right\}^{n+\nu}, \tag{9.38}$$

$$0<\nu<1, \quad 0\le \arg z\le \pi,$$

for if we apply (9.21) we obtain (9.33) after a change of variables. In fact, (9.38) is valid for $-\frac{1}{2}<\nu<1$, which can be shown from the general formula (A.7) in the Appendix.

If we multiply (9.38) by $e^{i\pi(\alpha+(1-\nu)/2)}$ and take the real part we obtain

$$\int_{-1}^{1}(1-t^2)^{\nu-1/2}C_n^{\nu}(t)k_{\alpha,\beta}(x-t)\,dt$$

$$=\frac{2^{1-\nu}\Gamma(n+2\nu)}{n!(n+\nu)\{\Gamma(\nu)\}^2}\cos\left\{(\nu+n)\sin^{-1}x-(n+\beta-\alpha)\frac{\pi}{2}\right\} \qquad (-1\le x\le 1) \tag{9.39}$$

where $\nu=\alpha+\beta+1, -\frac{1}{2}<\nu<1, -\frac{\pi}{2}\le \sin^{-1}x\le\frac{\pi}{2}$.

In connection with the second example, we have, from (5.19), (3.3), and (3.25),

$$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}P_n^{(\beta,\alpha)}(t)K_{\alpha,\beta}(z-t)\,dt$$

$$=\frac{(\nu)_n}{n!}\frac{1}{W_{\alpha,\beta}(z)}\int_{-1}^{1}\frac{(1-t)^{\alpha}(1+t)^{\beta}P_n^{(\alpha,\beta)}(t)\,dt}{\pi i(t-z)}, \tag{9.40}$$

$$-1<\alpha<0, \quad -1<\beta<0, \quad \nu=\alpha+\beta+1, \quad -1<\nu<1, \quad \operatorname{Im} z\ge 0.$$

On the real axis ($z=x+i0$), the integral on the right is interpreted as a Cauchy principal value for $-1<x<1$, and $K_{\alpha,\beta}(z)$ and $W_{\alpha,\beta}(z)$ are defined by (2.1), (2.2) and (3.13), (3.14), respectively.

Now Szegö [12, p. 95] gives the following representation of the Jacobi function of the second kind:

$$Q_n^{(\alpha,\beta)}(z)=\frac{1}{2}(z-1)^{-\alpha}(z+1)^{-\beta}\int_{-1}^{1}\frac{(1-t)^{\alpha}(1+t)^{\beta}}{z-t}P_n^{(\alpha,\beta)}(t)\,dt \tag{9.41}$$

where z is in the complex plane cut along the segment $(-1,1)$. The principal branch, single valued in the upper half plane, is taken to be real valued on the x axis for $x>1$. In our notation, then,

$$Q_n^{(\alpha,\beta)}(z)=\frac{1}{2}\frac{e^{i\pi\alpha}}{W_{\alpha,\beta}(z)}\int_{-1}^{1}\frac{(1-t)^{\alpha}(1+t)^{\beta}P_n^{(\alpha,\beta)}(t)}{z-t}dt, \qquad \operatorname{Im} z\ge 0. \tag{9.42}$$

Comparing (9.42) and (9.40) and recalling that $K_{\alpha,\beta}(z)=\frac{i}{\pi}e^{i\pi\alpha}z^{-\nu}$, we see that

$$\int_{-1}^{1}(1-t)^{\beta}(1+t)^{\alpha}P_n^{(\beta,\alpha)}(t)(z-t)^{-\nu}dt=\frac{2(\nu)_n}{n!}Q_n^{(\alpha,\beta)}(z), \tag{9.43}$$

$$\operatorname{Im} z\ge 0, \ -1<\alpha<0, \quad -1<\beta<0, \quad -1<\nu<1, \quad \nu=\alpha+\beta+1,$$

where on the real axis ($z=x+i0$) we take

$$(z-t)^{-\nu}=\begin{cases}|x-t|^{-\nu}, & x>t,\\ e^{-i\pi\nu}|x-t|^{-\nu}, & x<t.\end{cases}$$

Equation (9.43) appears to be a new integral representation of the Jacobi function of the second kind. At least it is not found in the standard reference works.

It is interesting to note that if a function $F(z)$ analytic in the upper half plane has the representation

$$F(z)=\frac{1}{\pi}\int_{-1}^{1}\frac{i}{z-t}f(t)\,dt \tag{9.44}$$

where

$$f(x)\begin{cases}\operatorname{Re}F(x)\geq 0 & \text{for } -1<x<1,\\ \equiv 0 & \text{for } x>1 \text{ and } x<-1\end{cases}$$

(and hence $F(z)$ is zero free in the upper half plane with $-\frac{\pi}{2}\leq\arg F(z)\leq\frac{\pi}{2}$), then with suitable smoothness conditions on $F(x)$, e.g., bounded variation, $\{F(z)\}^{\nu}$ has the representation

$$\{F(z)\}^{\nu}=\frac{1}{\pi}\int_{-1}^{1}\left\{\frac{i}{z-t}\right\}^{\nu}f_{\nu}(t)\,dt, \qquad 0<\nu\leq 1. \tag{9.45}$$

10. Kernel expansions. A formal expansion of a function $f(t)$, $-1<t<1$, in Jacobi polynomials $\{P_n^{(\lambda,\nu)}\}$ ($\lambda>-1$, $\mu>-1$), written

$$f(t)\sim\sum_{n=0}^{\infty}a_nP_n^{(\lambda,\mu)}(t) \tag{10.1}$$

is obtained from the orthogonality relation

$$\int_{-1}^{1}(1-t)^{\lambda}(1+t)^{\mu}P_n^{(\lambda,\mu)}(t)P_m^{(\lambda,\mu)}(t)\,dt \tag{10.2}$$

$$=\begin{cases}0 & (m\neq n)\\ \dfrac{\Gamma(\lambda+n+1)\Gamma(\mu+n+1)}{n!\Gamma(\lambda+\mu+n+1)}\cdot\dfrac{2^{\lambda+\mu+1}}{\lambda+\mu+2n+1} & (m=n).\end{cases}$$

So in (10.1) we have

$$a_n=\frac{\lambda+\mu+2n+1}{2^{\lambda+\mu+1}}\frac{n!\Gamma(\lambda+\mu+n+1)}{\Gamma(\lambda+n+1)\Gamma(\mu+n+1)}\int_{-1}^{1}(1-t)^{\lambda}(1+t)^{\mu}P_n^{(\lambda,\mu)}(t)f(t)\,dt, \tag{10.3}$$

assuming the last integral is convergent.

Szegö ([12, Chap. IX]) discusses the convergence of (10.1) and gives the following useful comparison theorem.

EQUICONVERGENCE THEOREM (SZEGÖ). *Let $f(t)$ be Lebesgue measurable in $[-1,1]$ and let the integrals*

$$\int_{-1}^{+1}(1-t)^{\lambda}(1+t)^{\mu}|f(t)|\,dt,$$

$$\int_{-1}^{+1}(1-t)^{\lambda/2-1/4}(1+t)^{\mu/2-1/4}|f(t)|\,dt$$

exist. If $S_n(t)$ *denotes the n-th partial sum of* (10.1) *and* $\sigma_n(\cos\theta)$ *the n-th partial sum of the Fourier (cosine) series of*

$$(1-\cos\theta)^{\lambda/2+1/4}(1+\cos\theta)^{\mu/2+1/4}f(\cos\theta),$$

then for $-1<t<1$

$$\lim_{n\to\infty}\left\{S_n(t)-(1-t)^{-\lambda/2-1/4}(1+t)^{-\mu/2-1/4}\sigma_n(t)\right\}=0$$

uniformly in $-1+\varepsilon\le t\le 1-\varepsilon$ *where* ε *is a fixed positive number,* $\varepsilon<1$.

For $f(t)=(x-t)^m k_{\alpha,\beta}(x-t)$ and $\lambda=\beta+r$, $\mu=\alpha+s$, where m,r, and s are integers, $m-\alpha-\beta>0$, formula (5.40) shows the coefficients a_n in (10.1) to be polynomials in x. We have

(10.4)

$$(x-t)^m k_{\alpha,\beta}(x-t)$$

$$\sim\frac{(-1)^{m+r}2^{m-\nu}}{\Gamma(\nu-m)}\sum_0^\infty\frac{(\lambda+\mu+2n+1)\Gamma(\lambda+\mu+n+1)\Gamma(\nu-m+n)}{\Gamma(\lambda+n+1)\Gamma(\mu+n+1)}P_n^{(\lambda,\mu)}(t)P_{n^*}^{(\alpha^*,\beta^*)}(x)$$

$$(-1<x<1)\quad(-1<t<1)$$

where $\lambda=\beta+r>-1$, $\mu=\alpha+s>-1$, $\nu=\alpha+\beta+1$, $m-\nu>-1$, $\alpha^*=\alpha-m-r$, $\beta^*=\beta-m-s$, $n^*=m+r+s+n\ge-1$ and $P_{-1}^{(\cdot)}\equiv 0$ (r,s,m are integers).

Now if $F(\theta)$ belongs to $L_1(-\pi,\pi)$, then $F_n(\theta)$, the nth partial sum of the Fourier series for $F(\theta)$, may be written

$$F_n(\theta)=\frac{1}{2\pi}\int_{-\pi}^{\pi}F(\varphi)\frac{\sin\{(n+1/2)(\theta-\varphi)\}}{\sin 1/2(\theta-\varphi)}d\varphi, \tag{10.5}$$

and then

$$F(\theta)-F_n(\theta)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{F(\theta)-F(\varphi)}{\sin(\theta-\varphi)/2}\sin\left\{\left(n+\frac{1}{2}\right)(\theta-\varphi)\right\}d\varphi. \tag{10.6}$$

Hence, if

$$\int_{-\pi}^{\pi}\left|\frac{F(\theta)-F(\varphi)}{\sin(\theta-\varphi)/2}\right|d\varphi<\infty, \tag{10.7}$$

then

$$\lim_{n\to\infty}\{F(\theta)-F_n(\theta)\}=0, \tag{10.8}$$

since, by the Riemann–Lebesgue lemma, the Fourier transform of a function of L_1 tends to zero at infinity.

Thus from Szegö's theorem we see that the partial sums of the series in (10.4) converge to the left-hand member (so we can replace $\sim$ by $=$) for $-1<x<1$, $-1<t<1$, $t\neq x$. Also the series will converge for $t=x$ ($x\neq\pm 1$) provided $m-\nu>0$.

In order to apply the theorem for $x=1$ we require

$$\int_{-1}^{1}(1-t)^{\lambda}(1+t)^{\mu}(1-t)^{m-\nu}dt<\infty \tag{10.9}$$

and

$$\int_{-1}^{1}(1-t)^{\lambda/2-1/4}(1-t)^{\mu/2-1/4}(1-t)^{m-\nu}dt<\infty, \tag{10.10}$$

i.e., since $\mu > -1$, we need

$$\lambda+m-\nu>1 \quad \text{and} \quad \frac{\lambda}{2}-\frac{1}{4}+m-\nu>-1. \tag{10.11}$$

So if (10.11) is satisfied the partial sums of the series (10.4) converge to $(x-t)^m k_{\alpha,\beta}(x-t)$ for $x=1$, $-1<t<1$. An analogous statement holds for $x=-1$ if λ in (10.11) is replaced by μ.

Further information about the nature of the convergence can be obtained from the asymptotic formula (cf. [4, p. 216])

$$P_n^{(a,b)}(\cos\theta)=\frac{\cos\left[\left(n+\frac{a+b+1}{2}\right)\theta-\frac{\pi}{4}(1+2a)\right]}{\sqrt{\pi n}\left(\sin\frac{\theta}{2}\right)^{a+1/2}\left(\cos\frac{\theta}{2}\right)^{b+1/2}}+O(n^{-3/2}) \tag{10.12}$$

$$(n\to\infty)$$

uniformly for $\varepsilon\le\theta\le\pi-\varepsilon$ $(0<\varepsilon<\pi)$ and

$$\frac{\Gamma(n+a)}{\Gamma(n+b)}=n^{a-b}\left[1+O(n^{-1})\right], \qquad n\to\infty. \tag{10.13}$$

Thus for the coefficient in (10.4) we have

(10.14)

$$\frac{(\lambda+\mu+2n+1)\Gamma(\lambda+\mu+n+1)\Gamma(\nu-m+n)}{\Gamma(\lambda+n+1)\Gamma(\mu+n+1)}=\frac{2}{n^{m-\nu}}\left\{1+O(n^{-1})\right\}, \qquad n\to\infty,$$

and hence for $-1<x<1$, $-1<t<1$, the series (10.4) converges (diverges) with

$$\sum_1^\infty \frac{A\cos n\varphi+B\sin n\varphi}{n^{m-\nu+1}}, \qquad \varphi=\cos^{-1}x-\cos^{-1}t, \tag{10.15}$$

so the series converges absolutely for $m-\nu>0$ ($-1<x<1$ and $-1<t<1$). Also we have

$$P_n^{(a,b)}(1)=\frac{(a+1)_n}{n!}=\frac{\Gamma(n+a+1)}{\Gamma(n+1)\Gamma(a+1)},$$

$$P_n^{(a,b)}(-1)=(-1)^n\frac{(b+1)_n}{n!}=\frac{\Gamma(n+b+1)}{\Gamma(n+1)\Gamma(b+1)}. \tag{10.16}$$

Now setting $x=1$ in (10.4) we have

$$P_{n^*}^{(\alpha^*,\beta^*)}(1)=\frac{\Gamma(n+\alpha+s+1)}{\Gamma(m+r+s+n+1)\Gamma(\alpha-m-r+1)}. \tag{10.17}$$

Then using (10.12) we find the series in (10.4) converges for $x=1$ and $-1<t<1$ provided only the second condition in (10.11) is satisfied. But what it converges to we are not sure in case the first condition in (10.11) is not satisfied. This difficulty arises when

$$\lim_{x\to1-}\int_{-1}^1(1-t)^\lambda(1+t)^\mu P_n^{(\lambda,\mu)}(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt$$

$$\neq\int_{-1}^1(1-t)^\lambda(1+t)^\mu P_n^{(\lambda,\mu)}(t)(1-t)^n k_{\alpha,\beta}(1-t)\,dt. \tag{10.18}$$

When the first condition in (10.11) is not satisfied, the integral on the right in (10.18) does not exist. However, the limit in (10.18) does exist (under the conditions given with (10.4)). The implication of (10.18) is that the kernel $x^m k_{\alpha,\beta}(x)$ is of the second type (changes sign) and $m-(\alpha+\beta+1)<0$.

We can obtain a rather simple expression for the partial sum of the series in (10.4) for $t=x$. Denoting the nth partial sum by $S_n^{(\lambda,\mu)}(t,x)$ we have

(10.19)
$$S_n^{(\lambda,\mu)}(t,x)=\int_{-1}^{1}(1-s)^{\lambda}(1+s)^{\mu}I_n^{(\lambda,\mu)}(t,s)(x-s)^m k_{\alpha,\beta}(x-s)\,ds \qquad (-1<x<1)$$

where the conditions in (10.4) are satisfied and

(10.20)
$$\begin{aligned} I_n^{(\lambda,\mu)}(t,s)&=\sum_{k=0}^{n}\{h_k(\lambda,\mu)\}^{-1}P_k^{(\lambda,\mu)}(t)P_k^{(\lambda,\mu)}(s)\\ &=C_n\frac{P_{n+1}^{(\lambda,\mu)}(t)P_n^{(\lambda,\mu)}(s)-P_n^{(\lambda,\mu)}(t)P_{n+1}^{(\lambda,\mu)}(s)}{t-s} \end{aligned}$$

and

$$h_k(\lambda,\mu)=\frac{2^{\lambda+\mu+1}}{2k+\lambda+\mu+1}\frac{\Gamma(k+\mu+1)\Gamma(k+\lambda+1)}{k!\Gamma(k+\lambda+\mu+1)},$$

$$C_n=\frac{2^{-\lambda-\mu}}{2n+\lambda+\mu+2}\frac{(n+1)!\Gamma(n+\lambda+\mu+2)}{\Gamma(n+\lambda+1)\Gamma(n+\mu+1)}.$$

Now in case $m-(\alpha+\beta+1)>0$, we have

(10.21)
$$\begin{aligned} S_n^{(\lambda,\mu)}(x,x)=C_n\int_{-1}^{1}(1-s)^{\lambda}(1+s)^{\mu}\{&P_{n+1}^{(\lambda,\mu)}(x)P_n^{(\lambda,\mu)}(s)\\ &-P_n^{(\lambda,\mu)}(x)P_{n+1}^{(\lambda,\mu)}(s)\}(x-s)^{m-1}k_{\alpha,\beta}(x-s)\,ds. \end{aligned}$$

Then from (5.40) it follows that

(10.22)
$$\begin{aligned} S_n^{(\lambda,\mu)}(x,x)=(-1)^{m+r+1}2^{m+r+s-1}C_n\Bigg\{&\frac{(\nu-m+1)_n}{n!}P_{n+1}^{(\lambda,\mu)}(x)P_{n^*-1}^{(\alpha^*+1,\beta^*+1)}(x)\\ &-\frac{(\nu-m+1)_{n+1}}{(n+1)!}P_n^{(\lambda,\mu)}(x)P_{n^*}^{(\alpha^*+1,\beta^*+1)}(x)\Bigg\} \end{aligned}$$

where the parameters are defined in (10.4) and (10.20). Equation (10.22) actually holds for $m-(\alpha+\beta+1)>-1$. To see this we multiply (10.19) by $(x-t)$ and write, in an abbreviated notation,

(10.23)
$$\int_{-1}^{1}\{\cdot\}(x-t)(x-s)^m k_{\alpha,\beta}(x-s)\,ds=(x-t)S_n^{(\lambda,\mu)}(t,x), \qquad -1<x<1.$$

Then we note from (10.10) and (5.40) that

(10.24)
$$\begin{aligned} &\int_{-1}^{1}\{\cdot\}(t-s)(x-s)^m k_{\alpha,\beta}(x-s)\,ds\\ &=(-1)^{m+r}2^{m+r+s}C_n\Bigg\{\frac{(\nu-m)_n}{n!}P_{n+1}^{(\lambda,\mu)}(t)P_{n^*}^{(\alpha^*,\beta^*)}(x)\\ &\qquad-\frac{(\nu-m)_{n+1}}{(n+1)!}P_n^{(\lambda,\mu)}(t)P_{n^*+1}^{(\alpha^*,\beta^*)}(x)\Bigg\}, \qquad -1<x<1. \end{aligned}$$

Adding (10.23) and (10.24) we obtain

$$\begin{aligned}&\int_{-1}^{1}\{\cdot\}(x-s)^{m+1}k_{\alpha,\beta}(x-s)\,ds\\&\quad=(x-t)S_n^{\lambda,\mu}(t,x)+(-1)^{m+r}2^{m+r+s}\\&\quad\cdot C_n\left\{\frac{(\nu-m)_n}{n!}P_{n+1}^{(\lambda,\mu)}(t)P_{n^*}^{(\alpha^*,\beta^*)}(x)\right.\\&\qquad\left.-\frac{(\nu-m)_{n+1}}{(n+1)!}P_n^{(\lambda,\nu)}(t)P_{n^*+1}^{(\alpha^*,\beta^*)}(x)\right\},\qquad -1<x<1.\end{aligned}\tag{10.25}$$

Then differentiating (10.25) with respect to x we have

$$\begin{aligned}(m+1-\nu)S_n^{\lambda,\mu}(t,x)&=\frac{\partial}{\partial x}\{(x-t)S_n^{\lambda,\mu}(t,x)\}\\&\quad+(-1)^{m+r}2^{m+r+s-1}C_n\left\{\frac{(\nu-m)_{n+1}}{n!}P_{n+1}^{(\lambda,\mu)}(t)P_{n^*-1}^{(\alpha^*+1,\beta^*+1)}(x)\right.\\&\qquad\left.-\frac{(\nu-m)_{n+2}}{(n+1)!}P_n^{(\lambda,\mu)}(t)P_{n^*}^{(\alpha^*+1,\beta^*+1)}(x)\right\}\end{aligned}\tag{10.26}$$

and hence

$$\begin{aligned}&(\nu-m)S_n^{(\lambda,\mu)}(t,x)+(x-t)\frac{\partial}{\partial x}S_n^{(\lambda,\mu)}(t,x)\\&\quad=(-1)^{m+r+1}2^{m+r+s-1}C_n\left\{\frac{(\nu-m)_{n+1}}{n!}P_{n+1}^{(\lambda,\mu)}(t)P_{n^*-1}^{(\alpha^*+1,\beta^*+1)}(x)\right.\\&\qquad\left.-\frac{(\nu-m)_{n+2}}{(n+1)!}P_n^{(\lambda,\mu)}(t)P_{n^*}^{(\alpha^*+1,\beta^*+1)}(x)\right\},\end{aligned}\tag{10.27}$$

which is valid for $m-\nu>-1$, and therefore (10.22) is valid for $m-\nu>-1$. As a special case ($m=r=s=0$) we have

$$\begin{aligned}&\sum_{k=0}^{h}\frac{(\nu)_k}{k!h_k(\alpha,\beta)}P_k^{(\alpha,\beta)}(x)P_k^{(\beta,\alpha)}(x)\\&\quad=\frac{2^{-\nu}}{2n+\nu+1}\frac{\Gamma(\nu+1)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\frac{(\nu+1)_n}{(\alpha+1)_n}\frac{(\nu+1)_n}{(\beta+1)_n}\\&\quad\cdot\{(\nu+n+1)P_n^{(\beta,\alpha)}(x)P_n^{(\alpha+1,\beta+1)}(x)-(n+1)P_{n+1}^{(\beta,\alpha)}(x)P_{n-1}^{(\alpha+1,\beta+1)}(x)\},\\&\qquad -1<\alpha<0,\quad -1<\beta<0,\quad \nu=\alpha+\beta+1.\end{aligned}\tag{10.28}$$

The polynomial on the right is evidently even, since $P_k^{(\alpha,\beta)}(-x)=(-1)^kP_k^{\beta,\alpha}(x)$; i.e., $P_k^{(\alpha,\beta)}(x)P_k^{(\beta,\alpha)}(x)$ is an even function of x.

Also, as a special case of (10.4), we have

$$k_{\alpha,\beta}(x-t)=\frac{2^{-\nu}\Gamma(\nu+1)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\sum_0^{\infty}\left(1+\frac{2n}{\nu}\right)\frac{(\nu)_n}{(\alpha+1)_n}\frac{(\nu)_n}{(\beta+1)_n}P_n^{(\beta,\alpha)}(t)P_n^{(\alpha,\beta)}(x),\tag{10.29}$$

where $-1<x<1$, $-1<t<1$, $-1<\alpha<0$, $-1<\beta<0$, $\nu=\alpha+\beta+1$ ($x\neq t$ unless $\nu<0$). The question of extending the validity of (10.29) is a matter for further study. It can be shown that it remains valid for α and β in the open interval $(-m-1,-m)$ where m is a positive integer.

Another interesting expansion in the Gegenbauer polynomials is obtained from (9.38). The orthogonality relation for these polynomials is

$$(10.30)\quad \int_{-1}^{1}(1-t^2)^{\nu-1/2}C_n^{\nu}(t)C_m^{\nu}(t)\,dt=\begin{cases}0 & (m\neq n),\\ \dfrac{\pi 2^{1-2\nu}\Gamma(n+2\nu)}{n!(\nu+n)\{\Gamma(\nu)\}^2} & (n=m).\end{cases}$$

We can rewrite (9.38) as

$$(10.31)\quad \int_{-1}^{1}(1-t^2)^{\nu-1/2}C_n^{\nu}(t)(z-t)^{-\nu}dt = \frac{\pi 2^{1-\nu}\Gamma(n+2\nu)}{n!(\nu+n)\{\Gamma(\nu)\}^2}\left(z-i\sqrt{1-z^2}\right)^{n+\nu},\qquad 0<\nu<1.$$

Here z may be any point in the complex plane cut along $(-1,1)$. The correct correspondence between the branches is established by taking on the real axis ($z=x+i0$)

$$(z-t)^{-\nu}=\begin{cases}|x-t|^{-\nu}, & x>t,\\ e^{-i\pi\nu}|x-t|^{-\nu}, & x<t,\end{cases}$$

$$\sqrt{1-z^2}=\begin{cases}|1-x^2|^{1/2}, & -1<x<1,\\ -i|1-x^2|^{1/2}, & x>1,\\ +i|1-x^2|^{1/2}, & x<-1,\end{cases}$$

$$-\pi\leq\arg\left(z-i\sqrt{1-z^2}\right)\leq 0,\qquad -1\leq x\leq 1,$$

$$\arg\left(z-i\sqrt{1-z^2}\right)=\begin{cases}0, & x>1,\\ -\pi, & x<-1.\end{cases}$$

Then with this correspondence of the branches we have

$$(10.32)\quad (z-t)^{-\nu}=2^{\nu}\sum_{n=0}^{\infty}C_n^{\nu}(t)\left(z-i\sqrt{1-z^2}\right)^{n+\nu},\qquad 0<\nu<1,\quad -1<t<1\quad (z\neq t),$$

and thus

$$(10.33)\quad 2^{\nu}\sum_{n=0}^{\infty}C_n^{\nu}(t)\cos\{(n+\nu)\cos^{-1}x\}=\begin{cases}|x-t|^{-\nu}, & -1<t<x<1,\\ (\cos\pi\nu)|x-t|^{\nu}, & -1<x<t<1\end{cases}$$

$$(0<\cos^{-1}x<\pi),\quad (0<\nu<1),$$

$$(10.34)\quad 2^{\nu}\sum_{n=0}^{\infty}C_n^{\nu}(t)\sin\{(n+\nu)\cos^{-1}x\}=\begin{cases}0, & -1<t<x<1,\\ (\sin\pi\nu)|x-t|^{-\nu}, & -1<x<t<1\end{cases}$$

$$(0<\cos^{-1}x<\pi)\quad (0<\nu<1),$$

$$(10.35)\qquad \frac{2^{\nu}}{\pi}\sum_{n=0}^{\infty} C_n^{\nu}(t)\sin\{(n+\nu)\cos^{-1}x-\pi\alpha\}=k_{\alpha,\beta}(x-t),$$

$$-1<x<1,\quad -1<t<1,\quad x\neq t\quad (0<\cos^{-1}x<\pi),\quad \nu=\alpha+\beta+1,\quad 0<\nu<1.$$

We note that (10.32) and hence (10.31) can be obtained from the generating function of the Gegenbauer polynomials:

$$(10.36)\qquad (1-2wt+w^2)^{-\nu}=\sum_{n=0}^{\infty} C_n^{\nu}(t)w^n,\qquad |w|<1,\quad \nu\neq 0.$$

It is a relatively simple matter to discuss the convergence and extend the validity of (10.32), since by a change of variables $z=\frac{1}{2}(w+w^{-1})$ we obtain (10.36).

Now the function $F_{\nu}(w;t)\equiv F_{\nu}(w)=(1-2wt+w^2)^{-\nu}$, where $-1\leq t\leq 1$, is analytic for $|w|<1$ and has singularities at $w=t\pm\sqrt{1-t^2}$. For $\nu<1$, $F_{\nu}(e^{i\theta})$ belongs to L_1 on the unit circle provided $-1<t<1$. It is easy to see that the partial sums given by

$$(10.37)\qquad \sum_{k=0}^{n} C_k^{\nu}(t)e^{ik\theta}=\frac{1}{2\pi}\int_{-1}^{1} F_{\nu}(e^{i\varphi})\frac{\sin\{(n+1/2)(\theta-\varphi)\}}{\sin(\theta-\varphi)/2}d\varphi$$

converge to $F_{\nu}(e^{i\theta})$ for $e^{i\theta}\neq t\pm i\sqrt{1-t^2}$, $-1<t<1$, $\nu<1$, with convergence also for $e^{i\theta}=t\pm i\sqrt{1-t^2}$, $-1\leq t\leq 1$, provided $\nu<0$. Now in case $t=\pm 1$, the singularities coalesce and we have to replace $\nu<1$ by $\nu<\frac{1}{2}$ in the preceding statement.

We should note that the condition $\nu\neq 0$ in (10.36) arises from the fact that C_k^{ν} is conventionally normalized differently for $\nu=0$ and so is not given by the generating function (10.36). However, if we agree to take $C_n^0(t)\equiv 0$ for $n>1$ and $C_0^0(t)=1$, then (10.32) is valid for $\nu<1$, the series converging for $-1<t<1$, $z\neq t$, and converges for $\nu<\frac{1}{2}$ in case $t=\pm 1$, $z\neq t$. The series converges absolutely for $\nu\leq 0$, $1\leq t\leq 1$ and $|z-i\sqrt{1-z^2}|<1$.

11. A problem of potential theory. The equation arising in potential problems

$$(11.1)\qquad -\frac{1}{\pi}\int_{-1}^{1} f(t)\log|x-t|\,dt=g(x),\qquad -1<x<1,$$

is usually solved in series form using the relation

$$(11.2)\qquad -\frac{1}{\pi}\int_{-1}^{1}(1-t^2)^{-1/2}T_n(t)\log|x-t|\,dt=\begin{cases}\dfrac{1}{n}T_n(x), & -1<x<1,\quad n\neq 0,\\ \log 2, & n=0,\end{cases}$$

where T_n is the Chebyshev polynomial of the first kind. The relation (11.2) can be derived from our formula

$$(11.3)\qquad \frac{1}{\pi}\cos\frac{\pi\nu}{2}\int_{-1}^{1}(1-t^2)^{\alpha}P_n^{(\alpha,\alpha)}(t)|x-t|^{-\nu}dt=\frac{(\nu)_n}{n!}P_n^{(\alpha,\alpha)}(x),\qquad -1<x<1$$

where $\nu=2\alpha+1$ by subtracting 1 from the kernel $|x-t|^{-\nu}$ and the corresponding quantity from the right-hand side (i.e., 0 for $n\geq 1$ and

$$\frac{1}{\sqrt{\pi}}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+3/2)}\cos\frac{\pi\nu}{2}\quad \text{for } n=0),$$

then dividing by ν and taking limits as $\nu\to 0$.

From (11.1) and (11.2) we obtain

$$\int_{-1}^{1}(1-x^2)^{-1/2}T_n(x)g(x)\,dx=\begin{cases}\frac{1}{n}\int_{-1}^{1}T_n(t)f(t)\,dt, & n\neq 0\\ (\log 2)\int_{-1}^{1}f(t)\,dt, & n\neq 0.\end{cases} \tag{11.4}$$

Equation (11.1) is also frequently solved by the "analytic function method" similar to that of §3. Here we wish to point out that the method of §6 may also be applied to solve (11.1).

It is convenient to work with

$$-\frac{1}{\pi}\int_{-1}^{1}f_0(t)\log(x-t)\,dt=g_0(x), \qquad -1<x<1, \tag{11.5}$$

where

$$f_0(t)=f(t)-a(1-t^2)^{-1/2}, \tag{11.6}$$

$$g_0(x)=g(x)-a\log 2, \qquad -1<x<1, \tag{11.7}$$

with the constant a chosen so that

$$\int_{-1}^{1}f_0(t)\,dt=0, \tag{11.8}$$

i.e., in accordance with (11.4),

$$a=\frac{1}{\pi}\int_{-1}^{1}f(t)\,dt=(\pi\log 2)^{-1}\int_{-1}^{1}(1-t^2)^{-1/2}g(t)\,dt. \tag{11.9}$$

We have

$$-\frac{1}{\pi}\int_{-1}^{1}(1-t^2)^{-1/2}\log|x-t|\,dt\equiv s(x), \qquad -\infty<x<\infty. \tag{11.10}$$

Clearly, $s(x)$ is continuous (comparable to a convolution of functions in complementary L_p spaces) and according to (11.2)

$$s(x)=\log 2, \qquad -1<x<1. \tag{11.11}$$

Then

$$s'(x)=0, \qquad -1<x<1, \tag{11.12}$$

and

$$s'(x)=-\frac{1}{\pi}\int_{-1}^{1}(1-t^2)^{-1/2}\frac{dt}{x-t}=\begin{cases}-(x^2-1)^{-1/2}, & x>1,\\ (x^2-1)^{-1/2}, & x<-1.\end{cases} \tag{11.13}$$

Now we assume that f_0 in (11.5) belongs to L_1 so that the projection of g_0 on $(-1,1)$ belongs to L_p for every finite positive p. Thus for $-1<\tau<1$ we may define $\varphi(\tau)$ by

$$\varphi(\tau)=\int_{\tau}^{1}(x-\tau)^{-1/2}(1-x)^{-1/2}g_0(x)\,dx. \tag{11.14}$$

Then we may replace g_0 by the integral in (11.5) and interchange the order of integration to obtain

$$\varphi(\tau)=\int_{-1}^{1}K(\tau,t)f_0(t)\,dt \tag{11.15}$$

where

$$K(\tau,t)=-\frac{1}{\pi}\int_{\tau}^{1}(x-\tau)^{-1/2}(1-x)^{-1/2}\log|x-t|\,dx$$

$$=-\frac{1}{\pi}\int_{-1}^{1}(1-u^2)^{-1/2}\log\left|\frac{1-\tau}{2}u+\frac{1+\tau}{2}-t\right|du \tag{11.16}$$

$$=\log\frac{2}{1-\tau}+s\left(\frac{2t-1-\tau}{1-\tau}\right).$$

Then if we set $f_0(t)=\frac{d}{dt}\int_{-1}^{t}f_0(u)\,du$ in (11.15) and integrate by parts we obtain

$$\varphi(\tau)=-\int_{-1}^{\tau}\frac{dt}{\sqrt{\tau-t}}\int_{-1}^{t}f_0(u)\,dt, \tag{11.17}$$

and hence

$$\int_{-1}^{x}f_0(t)\,dt=-\frac{\sqrt{1-x}}{\pi}\frac{d}{dx}\int_{-1}^{x}\frac{\varphi(t)\,dt}{\sqrt{x-t}}. \tag{11.18}$$

Thus we obtain the solution to (11.5).

We may write, by a change of variables,

$$\varphi(t)=\int_{0}^{1}u^{-1/2}(1-u)^{-1/2}g_0\{(1-t)u+t\}\,du, \tag{11.19}$$

and then if g is differentiable,

$$\varphi'(t)=\int_{0}^{1}u^{-1/2}(1-u)^{1/2}g'\{(1-t)u+t\}\,du$$

$$=\frac{1}{1-t}\int_{t}^{1}\frac{\sqrt{1-x}}{\sqrt{x-t}}g'(x)\,dx, \tag{11.20}$$

and hence, since $\varphi(-1)=0$,

$$\int_{-1}^{x}f_0(t)\,dt=-\frac{\sqrt{1-x}}{\pi}\int_{-1}^{x}\frac{\varphi'(t)\,dt}{\sqrt{x-t}}. \tag{11.21}$$

Thus if g is differentiable

$$-\frac{1}{\pi}\int_{-1}^{1}\frac{f_0(t)}{x-t}dt=g'(x),\qquad -1<x<1, \tag{11.22}$$

and we know the solution ($\int f_0=0$) if it exists, is given by

$$f_0(x)=-\frac{1}{\pi\sqrt{1-x^2}}\int_{-1}^{1}\frac{\sqrt{1-t^2}\,g'(t)\,dt}{x-t}. \tag{11.23}$$

Hence, comparing (11.21) and (11.23), we have the identity for the finite Hilbert transform

(11.24)

$$\frac{1}{\sqrt{1-x^2}}\int_{-1}^{1}\frac{\sqrt{1-t^2}\,g'(t)\,dt}{x-t}=\frac{d}{dx}\left\{\sqrt{1-x}\int_{-1}^{1}\frac{dt}{\sqrt{x-t}}\frac{1}{1-t}\int_{t}^{1}\frac{\sqrt{1-u}}{\sqrt{u-t}}g'(u)\,du\right\}.$$

We can verify (11.18) by showing that it holds for

$$f_0(x)=T_n(x)(1-x^2)^{-1/2}, \qquad g_0(x)=\frac{1}{n}T_n(x), \qquad -1<x<1, \quad n=1,2,\cdots.$$

We have

$$T_n(x)=2^{2n}\frac{(n!)^2}{(2n)!}P_n^{(-1/2,-1/2)}(x), \tag{11.25}$$

$$\frac{d}{dx}T_n(x)=nU_{n-1}(x)=\frac{n}{2}2^{2n}\frac{(n!)^2}{(2n)!}P_{n-1}^{(1/2,1/2)}(x). \tag{11.26}$$

So from (6.31) we have for $g'(x)=U_{n-1}(x)$, $-1<x<1$,

$$\varphi'(t)=\frac{1}{1-t}\int_t^1\frac{\sqrt{1-x}}{\sqrt{x-t}}U_{n-1}(x)\,dx=\frac{\pi}{2}P_{n-1}^{(1,0)}(t), \tag{11.27}$$

and from (6.32)

$$\begin{aligned}\int_{-1}^{x}P_{n-1}^{(1,0)}(t)\frac{dt}{\sqrt{x-t}}&=\sqrt{1+x}\,\frac{\sqrt{\pi}\,(n-1)!}{\Gamma(n+\frac{1}{2})}P_{n-1}^{(1/2,1/2)}(x)\\&=\frac{2}{n}\sqrt{1+x}\,U_{n-1}(x).\end{aligned} \tag{11.28}$$

Then

$$\int_{-1}^{x}f_0(t)\,dt=-\frac{1}{n}\sqrt{1-x^2}\,U_{n-1}(x)=-\frac{1}{n}\sin(n\cos^{-1}x) \qquad (0\le\cos^{-1}x\le\pi). \tag{11.29}$$

Therefore

$$f_0(x)=\frac{\cos(n\cos^{-1}x)}{\sqrt{1-x^2}}=\frac{T_n(x)}{\sqrt{1-x^2}}, \qquad -1<x<1, \tag{11.30}$$

which agrees with (11.2).

Appendix. Here we derive equation (5.40) by evaluating the integral

$$\int_{-\infty}^{\infty}f(t)k_{\alpha,\beta}(x-t)\,dt=g(x), \tag{A.1}$$

for $-1<x<1$, where

$$f(t)=\begin{cases}(1+t)^{\lambda}(1-t)^{\mu}, & -1<t<1,\\ 0, & \text{otherwise.}\end{cases} \tag{A.2}$$

For λ and μ sufficiently large the nth derivative of f will belong to L_1 and we have

$$\begin{aligned}\left(\frac{d}{dx}\right)^n g(x)&=\left(\frac{d}{dx}\right)^n\int_{-\infty}^{\infty}k_{\alpha,\beta}(t)f(x-t)\,dt\\&=\int_{-\infty}^{\infty}f^{(n)}(t)k_{\alpha,\beta}(x-t)\,dt,\end{aligned} \tag{A.3}$$

where

$$f^{(n)}(t)=(-2)^n m!(1-t)^{\mu^*}(1+t)^{\lambda^*}P_m^{(\mu^*,\lambda^*)}(t), \tag{A.4}$$
$$-1<t<1, \quad \mu^*=\mu-n>-1, \quad \lambda^*=\lambda-n>-1.$$

We recall that

$$k_{\alpha,\beta}(x)=\begin{cases} -\frac{1}{\pi}\sin\pi\alpha|x|^{-\nu}, & x>0 \\ -\frac{1}{\pi}\sin\pi\beta|x|^{-\nu}, & x<0 \end{cases} \qquad (\nu=\alpha+\beta+1), \tag{A.5}$$

and here the only restriction on α and β is $\alpha+\beta<0$. We can express the integral in (A.1) in terms of hypergeometric functions using

$$\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}dt=\frac{\Gamma(b)\Gamma(c-b)}{\Gamma(c)}F(a,b;c;x). \tag{A.6}$$

We have

(A.7)

$$g(x)=-\frac{1}{\pi}\sin\pi\alpha\int_{-1}^{x}(1+t)^{\lambda}(1-t)^{\mu}(x-t)^{-\nu}dt-\frac{1}{\pi}\sin\pi\beta\int_{x}^{1}(1+t)^{\lambda}(1-t)^{\mu}(t-x)^{-\nu}dt$$
$$=-\frac{1}{\pi}\sin\pi\alpha(1+x)^{\lambda+1-\nu}2^{\mu}\frac{\Gamma(\lambda+1)\Gamma(1-\nu)}{\Gamma(\lambda+2-\nu)}F\left(-\mu,\lambda+1;\lambda+2-\nu;\frac{1+x}{2}\right)$$
$$-\frac{1}{\pi}\sin\pi\beta(1-x)^{\mu+1-\nu}2^{\lambda}\frac{\Gamma(\mu+1)\Gamma(1-\nu)}{\Gamma(\mu+2-\nu)}F\left(-\lambda,\mu+1;\mu+2-\nu;\frac{1-x}{2}\right),$$
$$\lambda>-1, \quad \mu>-1, \quad \nu=\alpha+\beta+1<1 \quad (-1<x<1).$$

Now we use the identity

(A.8)

$$F(a,b;c;x)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-b)\Gamma(c-a)}F(a,b;a+b-c+1;1-x)$$
$$+(1-x)^{c-a-b}\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}F(c-a,c-b;c-a-b+1;1-x),$$
$$0<x<1 \quad (a+b-c \text{ not an integer})$$

to write

(A.9)

$$F\left(-\mu,\lambda+1;\lambda+2-\nu;\frac{1+x}{2}\right)$$
$$=\frac{\Gamma(\lambda+2-\nu)\Gamma(1-\nu+\mu)}{\Gamma(\lambda+\mu+2-\nu)\Gamma(1-\nu)}F\left(-\mu,\lambda+1;-\mu+\nu;\frac{1-x}{2}\right)$$
$$+\left(\frac{1-x}{2}\right)^{\mu-\nu+1}\frac{\Gamma(\lambda+2-\nu)\Gamma(\nu-\mu-1)}{\Gamma(-\mu)\Gamma(\lambda+1)}F\left(\lambda+\mu+2-\nu,1-\nu;2-\nu+\mu;\frac{1-x}{2}\right),$$
$$\nu-\mu\neq \text{ integer}.$$

Then applying the identity

$$F(a,b;c;x)=(1-x)^{c-a-b}F(c-a,c-b;cx) \tag{A.10}$$

to (A.9), we obtain

(A.11)

$$F\left(-\mu,\lambda+1;\lambda+2-\nu;\frac{1+x}{2}\right)$$
$$=\left(\frac{1+x}{2}\right)^{\nu-\lambda-1}\frac{\Gamma(\lambda+2-\nu)\Gamma(1-\nu+\mu)}{\Gamma(\lambda+\mu+2-\nu)\Gamma(1-\nu)}F\left(\nu,\nu-\mu-\lambda-1;-\mu+\nu;\frac{1-x}{2}\right)$$
$$+\left(\frac{1-x}{2}\right)^{\mu-\nu+1}\left(\frac{1+x}{2}\right)^{\nu-\lambda-1}\frac{\Gamma(\lambda+2-\nu)\Gamma(\nu-\mu-1)}{\Gamma(-\mu)\Gamma(\lambda+1)}$$
$$\cdot F\left(-\lambda,\mu+1;\mu-\nu+2;\frac{1-x}{2}\right),\qquad \nu-\mu\neq \text{integer}.$$

Then substituting (A.11) in (A.7) we have

(A.12)

$$g(x)=-\frac{1}{\pi}\sin\pi\alpha 2^{\mu+\lambda+1-\nu}\frac{\Gamma(1-\nu+\mu)\Gamma(\lambda+1)}{\Gamma(\lambda+\mu+2-\nu)}F\left(\nu,\nu-\mu-\lambda-1;-\mu+\nu;\frac{1-x}{2}\right)$$
$$-\frac{1}{\pi}\sin\pi\alpha(1-x)^{\mu-\nu+1}2^{\lambda}\frac{\Gamma(1-\nu)\Gamma(\nu-\mu-1)}{\Gamma(-\mu)}F\left(-\lambda,\mu+1;\mu-\nu+2;\frac{1-x}{2}\right)$$
$$-\frac{1}{\pi}\sin\pi\beta(1-x)^{\mu-\nu+1}2^{\lambda}\frac{\Gamma(1+\mu)\Gamma(1-\nu)}{\Gamma(\mu+2-\nu)}F\left(-\lambda,\mu+1;\mu-\nu+2;\frac{1-x}{2}\right)$$
$$\nu-\mu\neq \text{integer}.$$

In order to combine the last two terms in (A.12) we use the identity $\Gamma(z)\Gamma(1-z)=\pi/\sin\pi z$. Thus

(A.13)
$$\frac{1}{\Gamma(-\mu)}=-\Gamma(1+\mu)\frac{\sin\pi\mu}{\pi},$$

(A.14)
$$\Gamma(\nu-\mu-1)=-\frac{\pi}{\sin\pi(\nu-\mu)}\frac{1}{\Gamma(\mu+2-\nu)}$$

or

(A.15)
$$\frac{\Gamma(\nu-\mu-1)}{\Gamma(-\mu)}=\frac{\sin\pi\mu}{\sin\pi(\nu-\mu)}\frac{\Gamma(1+\mu)}{\Gamma(\mu+2-\nu)}.$$

Using (A.15) in (A.12) and combining the last two terms we obtain the factor

(A.16)
$$-\frac{\sin\pi\alpha\sin\pi\mu}{\sin\pi(\nu-\mu)}-\sin\pi\beta=\frac{\sin\nu\pi\sin(\mu-\beta)\pi}{\sin\pi(\nu-\mu)},$$

and hence

(A.17)

$$g(x)=-\frac{1}{\pi}\sin\pi\alpha 2^{\mu+\lambda+1-\nu}\frac{\Gamma(1-\nu+\mu)\Gamma(\lambda+1)}{\Gamma(\lambda+\mu+2-\nu)}F\left(\nu,\nu-\mu-\lambda-1;-\mu+\nu;\frac{1-x}{2}\right)$$
$$+\frac{\sin\nu\pi\sin(\mu-\beta)\pi}{\pi\sin\pi(\nu-\mu)}(1-x)^{\mu-\nu+1}2^{\lambda}\frac{\Gamma(1+\mu)\Gamma(1-\nu)}{\Gamma(\mu+2-\nu)}$$
$$\cdot F\left(-\lambda,\mu+1;\mu-\nu+2;\frac{1-x}{2}\right),\qquad \nu-\mu\neq \text{integer}.$$

Then setting $\mu=\beta+r$, $\lambda=\alpha+s$, where r and s are integers and $\mu-\nu\neq$ integer (i.e., $\alpha\neq$ integer) we have, noting that $F(a,b;c;x)=F(b,a;c;x)$,

$$\text{(A.18)}\qquad \begin{aligned}&\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}k_{\alpha,\beta}(x-t)\,dt\\&\quad=-\frac{1}{\pi}\sin\pi\alpha\,2^{r+s}\frac{\Gamma(r-\alpha)\Gamma(\alpha+s+1)}{\Gamma(r+s+1)}F\left(-r-s,\nu;\alpha+1-r;\frac{1-x}{2}\right)\end{aligned}$$

where r and s are integers, $\beta+r>-1$, $\alpha+s>-1$, $\alpha+\beta<0$ (hence $r+s\geq-1$), $\alpha\neq$ integer, $\nu=\alpha+\beta+1$ $(-1<x<1)$.

We note that in case $r+s=-1$, the right-hand side of (A.18) vanishes $(-1<x<1)$.

Now

$$\text{(A.19)}\qquad P_n^{(\alpha,\beta)}(x)=\frac{\Gamma(n+a+1)}{n!\Gamma(a+1)}F\left(-n,a+b+n+1;a+1;\frac{1-x}{2}\right)$$

and

$$\text{(A.20)}\qquad \Gamma(r-\alpha)\Gamma(1+\alpha-r)=\frac{\pi}{\sin\pi(r-\alpha)}=(-1)^{r+1}\frac{\pi}{\sin\pi\alpha}.$$

Hence,

(A.21)

$$\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}k_{\alpha,\beta}(x-t)\,dt=(-1)^r2^{r+s}P_{r+s}^{(\alpha-r,\beta-s)}(x)\qquad(-1<x<1),$$

where r and s are integers, $\beta+r>-1$, $\alpha+s>-1$, $\alpha+\beta<0$ (hence $r+s\geq-1$), $P_{-1}^{(\cdot)}\equiv0$, $\alpha\neq$ integer. Now in case α is an integer, we have from (A.7)

$$\text{(A.22)}\qquad \begin{aligned}&\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}k_{\alpha,\beta}(x-t)\,dt\\&\quad=-\frac{1}{\pi}\sin\pi\beta(1-x)^{r-\alpha}2^{\alpha+s}\frac{\Gamma(\beta+r+1)\Gamma(1-\nu)}{\Gamma(r-\alpha+1)}\\&\qquad\cdot F\left(-\alpha-s,\beta+r+1;r-\alpha+1;\frac{1-x}{2}\right)\qquad(-1<x<1)\end{aligned}$$

where $r,s,$ and α are integers, $\beta+r>-1$, $\alpha+s\geq0$, $\alpha+\beta<0$ (hence $r+s\geq0$), $\nu=\alpha+\beta+1$. Since $r>-1-\beta$, and $\alpha+\beta<0$, it follows from the fact that r and α are integers that $r-\alpha\geq0$. So the expression on the right in (A.22) is a polynomial of degree $(r+s)$ in x. From (A.19) we have

$$\text{(A.23)}\qquad F\left(-\alpha-s,\beta+r+1;r-\alpha+1;\frac{1-x}{2}\right)=\frac{(\alpha+s)!\Gamma(r-\alpha+1)}{(r+s)!}P_{\alpha+s}^{(r-\alpha,\beta-s)}(x).$$

Then we can write (A.22) as

$$\text{(A.24)}\qquad \begin{aligned}&\int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}k_{\alpha,\beta}(x-t)\,dt\\&\quad=-\frac{1}{\pi}\sin\pi\beta\left(\frac{1-x}{2}\right)^{r-\alpha}2^{r+s}\Gamma(\beta+r+1)\frac{\Gamma(1-\nu)(\alpha+s)!}{(r+s)!}P_{\alpha+s}^{(r-\alpha,\beta-s)}(x),\\&\qquad\qquad-1<x<1,\end{aligned}$$

with the conditions in (A.22). Now if $r=\alpha$, we have

$$-\frac{1}{\pi}\sin\pi\beta\Gamma(\beta+r+1)\Gamma(1-\nu)=-\frac{1}{\pi}\sin\pi(\nu-1-r)\Gamma(\nu)\Gamma(1-\nu)=(-1)^r.$$

So (A.21) is valid for $\alpha=r$. In case $r-\alpha\geq 1$ (r and α integers), we have (cf. Szegö [3, his formula (4.22.2)])

$$\text{(A.25)}\qquad \frac{\Gamma(r+\beta+1)}{\Gamma(\alpha+\beta+1)}\left(\frac{x-1}{2}\right)^{r-\alpha}P_{\alpha+s}^{(r-\alpha,\beta-s)}(x)=\frac{(r+s)!}{(s+\alpha)!}P_{r+s}^{(\alpha-r,\beta-s)}(x).$$

We find then that (A.23) agrees with (A.21) when α is an integer. Therefore:

(A.26) Equation (A.21) is valid with the condition ($\alpha\neq$ integer) removed.

Now, if in (A.21) we replace β by $(\beta-m)$ and r by $(r+m)$, where m is an integer, using the fact (cf. (2.25)) $k_{\alpha,\beta-m}(x)=x^m k_{\alpha,\beta}(x)$, we obtain

$$\text{(A.27)}\qquad \int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}(x-t)^m k_{\alpha,\beta}(x-t)\,dt$$
$$=(-1)^{r+m}2^{r+s+m}P_{m+r+s}^{(a,b)}(x)\qquad(-1<x<1),$$

where r,s and m are integers, $\beta+r>-1$, $\alpha+s>-1$, $\alpha+\beta<m$ (hence $m+r+s\geq -1$), $P_{-1}^{(\cdot)}\equiv 0$, $a=\alpha-r-m$, $b=\beta-s-m$. Now if in (A.27) we suppose that

$$\beta+r>n-1,\qquad \alpha+s>n-1,$$

where n is some nonnegative integer, we can apply (A.3) and (A.4) together with

$$\text{(A.28)}\qquad \left(\frac{d}{dx}\right)^n P_k^{(a,b)}(x)=2^{-n}(a+b+k+1)_n P_{k-n}^{(a+n,b+n)}(x)$$

to obtain

$$\text{(A.29)}\qquad (-2)^n n!\int_{-1}^{1}(1-t)^c(1+t)^d P_n^{(c,d)}(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt$$
$$=(-1)^{r+m}2^{r+s+m-n}(\alpha+\beta+1-m)_n P_{m+r+s-n}^{(p,q)}(x)\qquad(-1<x<1)$$

where $c=\beta+r-n>-1$, $d=\alpha+s-n>-1$, $p=\alpha-r-m+n$, $q=\beta-s-m+n$, $\alpha+\beta<m$, and r,s,m,n are integers ($n\geq 0$) (hence $m+r+s-n\geq n-1$) and $P_{-1}^{(\cdot)}\equiv 0$. Finally, replacing r by $(r+n)$ and s by $(s+n)$, we have

$$\text{(A.30)}\qquad \int_{-1}^{1}(1-t)^{\beta+r}(1+t)^{\alpha+s}P_n^{(\beta+r,\alpha+s)}(t)(x-t)^m k_{\alpha,\beta}(x-t)\,dt$$
$$=(-1)^{m+r}2^{m+r+s}\frac{(\alpha+\beta+1-m)_n}{n!}P_{n^*}^{(\alpha^*,\beta^*)}(x)\qquad(-1<x<1)$$

where $\beta+r>-1$, $\alpha+s>-1$, $\alpha+\beta<m$, $n^*=m+r+s+n$, $\alpha^*=\alpha-m-r$, $\beta^*=\beta-m-s$, and r,s,m,n are integers ($n\geq 0$) (hence $n^*\geq n-1$) and $P_{-1}^{(\cdot)}\equiv 0$.

Acknowledgment. The author would like to thank J. C. Lagarias for help in preparing the manuscript, and R. Askey for pointing out the relevance of [6].

REFERENCES

[1] T. Carleman (1922), *Über die Abelsche Integralgleichung mit konstenten Integrationsgrenge*, Math. Z., 15, pp. 111–120.

[2] J. A. Cochran (1972), *The Analysis of Linear Integral Equations*, McGraw-Hill, New York.

[3] G. E. Latta (1956), *The solution of a class of integral equations*, J. Rational Mech. Anal., 5, pp. 821–834.

[4] W. Magnus, F. Oberhettinger and R. Soni (1966), *Formulas and Theorems for the Special Functions of Mathematical Physics*, 3rd enlarged ed., Springer-Verlag, New York.

[5] C. E. PEARSON (1957/58), *On the finite strip problem*, Quart. Applied Math., 15, pp. 203–208.
[6] G. PÓLYA AND G. SZEGÖ (1931), *Über den transfiniten Durchmesser (Kapazitätskonstant) von ebenen und räumlichen Punktmengen*, J. Reine Angew. Math., 165, pp. 4–49.
[7] M. RAHMAN (1976), *Construction of a family of positive kernels from Jacobi polynomials*, this Journal, 7, pp. 92–116.
[8] _____ (1976), *A five-parameter family of positive kernels from Jacobi polynomials*, this Journal, 7, pp. 386–413.
[9] M. SHINBROT (1958/59), *A generalization of Latta's method for the solution of integral equations*, Quart. Applied Math., 16, pp. 415–421.
[10] _____ (1970/71), *The solution of some integral equations of Wiener–Hopf type*, Quart. Applied Math., 28, pp. 15–36.
[11] I. STAKGOLD (1968), *Boundary Value Problems of Mathematical Physics*, Vol. II, Macmillan, New York, p. 191.
[12] G. SZEGÖ (1959), *Orthogonal Polynomials*, AMS Colloquium Publications, Vol. XXIII, American Mathematical Society, New York.
[13] E. C. TITCHMARSH (1948), *Introduction to the Theory of the Fourier Integral*, 2nd ed., Oxford Univ. Press.
[14] F. G. TRICOMI (1951), *On the finite Hilbert transform*, Quart. J. Math. (Oxford), 2, pp. 199–211.
[15] _____ (1957), *Integral Equations*, Interscience, New York, esp. pp. 179–185.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0009 $01.25/0

δ-FRACTION EXPANSIONS OF ANALYTIC FUNCTIONS*

L. J. LANGE†

Abstract. In this paper a *δ-fraction* is defined to be a finite or infinite continued fraction K of the form $\mathsf{K}=b_0-\delta_0 z+\mathsf{K}_{n=1}^{\infty}(d_n z/(1-\delta_n z))$, where $b_0, d_n\in\mathbb{C}$ (complex numbers), $d_n\neq 0$, and δ_n is either 0 or 1 for each n. K is said to be *regular* if $d_{k+1}=1$ for each k such that $\delta_k=1$. An extensive investigation of these δ-fractions is made from the points of view of their correspondence to power series, their convergence properties and their capacity to represent analytic functions. Five basic theorems are given dealing with the correspondence between δ-fractions and power series of the form $L_0=1+c_1z+c_2z^2+\cdots$. It is shown, for example, that (a) there is a unique regular δ-fraction corresponding to each power series L_0, and (b) L_0 represents a rational function in a neighborhood of $z=0$, if and only if, its corresponding δ-fraction terminates. Seven convergence theorems involving δ-fractions are given. The first such theorem, which ties in Poincare's theorem on finite difference equations with continued fractions, gives information on the convergence behavior of the sequence $\{B_n(z)/B_{n-1}(z)\}$ of ratios of approximant denominators of K when both sequences $\{d_n\}$ and $\{\delta_n\}$ converge. Three of the convergence theorems are concerned with what are called (p,q) *limit periodic δ-fractions* of types (1,1) (1,2) and (2,1). Many explicit δ-fraction expansions are given for a variety of classical analytic functions and asymptotic series, along with considerable information about the regions of validity of these representations. For example, it is shown that

$$\tan z=\cfrac{z}{1-z+\cfrac{z}{1+\cfrac{z/3}{1-\cfrac{z/3}{1-\cfrac{z/5}{1+\cfrac{z/5}{1+\cfrac{z/7}{1-\cfrac{z/7}{1-\cfrac{z/9}{1+\cdots}}}}}}}}}$$

is valid for all $z\in\mathbb{C}$. A method of successive extensions is used in obtaining these expansions, and Poincaré's theorem plays an important role in establishing the regions of validity.

Introduction. By a *δ-fraction* we mean a finite or infinite continued fraction of the form

$$b_0-\delta_0 z+\cfrac{d_1z}{1-\delta_1z}+\cfrac{d_2z}{1-\delta_2z}+\cfrac{d_3z}{1-\delta_3z}+\cdots, \tag{1.1}$$

where b_0 and the d_n are complex constants, $d_n\neq 0$ for $n\geq 1$, and the δ_n are real constants restricted to the values 0 or 1. We adopt the convention that the δ-fraction (1.1), and all of its approximants, have value b_0 at $z=0$. We say that the δ-fraction (1.1) is *regular* if $d_{k+1}=1$ for each k such that $\delta_k=1$. We choose the name δ-fraction for the continued fraction (1.1) because of the binary "impulse" nature of the sequence $\{\delta_n\}$ and the analogies, therefore, with the δ's in the Dirac delta function and the Kronecker delta symbol. We are led to this investigation of δ-fractions, and their connections with analytic functions, in our quest to find an answer to the following question: Is there a class $\mathfrak{D}$ of "simple" continued fractions

$$b_0(z)+\cfrac{a_1(z)}{b_1(z)}+\cfrac{a_2(z)}{b_2(z)}+\cfrac{a_3(z)}{b_3(z)}+\cdots \tag{1.2}$$

having the following desirable properties?

(a) The elements $a_n(z)$ and $b_n(z)$ are polynomials in z of degree ≤ 1.

(b) $\mathfrak{D}$ contains the class of regular C-fractions

$$c_0+\cfrac{c_1z}{1}+\cfrac{c_2z}{1}+\cfrac{c_3z}{1}+\cdots, \qquad c_n\in\mathbb{C},\quad c_n\neq 0 \text{ if } n\geq 1.$$

* Received by the editors June 22, 1981. This research was funded in part by a grant from the Research Council of the Graduate School, University of Missouri-Columbia.

† Department of Mathematics, University of Missouri, Columbia, Missouri, 65211.

(c) Given a power series

$$L_0 \equiv c_0 + c_1 z + c_2 z^2 + c_3 z^3 + \cdots, \qquad c_n \in \mathbb{C}, \quad c_0 \neq 0,$$

there exists a unique member $\mathsf{K}_0 \in \mathfrak{D}$ such that K_0 *corresponds* to L_0, i.e., the Maclaurin series of the nth approximant of K_0 agrees termwise with L_0 up to and including the term $c_{k(n)} z^{k(n)}$, where $k(n) \to \infty$ as $n \to \infty$.

(d) If L_0 represents a rational function in a neighborhood of $z=0$, then its corresponding $\mathsf{K}_0 \in \mathfrak{D}$ terminates.

(e) Let us say that $\mathsf{K} \in \mathfrak{D}$ corresponds to a function $f(z)$, analytic at $z=0$, if K corresponds to the Maclaurin series of f. Then for many classical functions, analytic in a neighborhood of the origin, explicit and useful formulas can be obtained for the elements $a_n(z)$ and $b_n(z)$ of the continued fractions in $\mathfrak{D}$ corresponding to these functions.

(f) Much information can be given about the convergence of the continued fractions in $\mathfrak{D}$ that correspond to functions which are analytic at the origin.

(g) In many cases, at least "some" of the approximants of the continued fraction $\mathsf{K} \in \mathfrak{D}$ corresponding to a power series L_0 are in the Padé table for L_0.

The C-fractions of Leighton and Scott [13], the T-fractions of Thron [26], and the P-fractions of Magnus [14], [15] all essentially meet requirement (c), but each of these classes fails to meet one or more of the remaining requirements. The C-fractions in general do not meet requirement (a), and the regular C-fractions above do not meet requirement (c). For example, it is known that there is no regular C-fraction corresponding to $1+z^2$. The T-fractions essentially meet requirements (a) and (f), in addition to (c), but they do not meet (b), (d) and (g). The P-fractions have a close connection with the Padé table of a given power series L_0, but they fail to meet requirements (a) and (b), among others. The class of general T-fractions, studied by Waadeland [31], [32] and others, essentially contains our class of δ-fractions, but, to date, this general class has been studied more from the interpolation point of view, i.e., more from the point of view of their connections with two-point Padé tables for meromorphic functions. Closely related to the general T-fractions are the M-fractions studied by a number of authors [2], [17], [18]. For thorough treatments of the subject of correspondence and the properties of many kinds of corresponding continued fractions, including the ones just mentioned, we refer the reader to the excellent, new and up-to-date book on continued fractions by Jones and Thron [9].

We offer the class of regular δ-fractions as our candidate for an answer to the question posed above. It is easily seen that these continued fractions satisfy (a) and (b). In §2 we give five basic theorems dealing with the correspondence between δ-fractions and power series. It follows from Theorems 2.1 and 2.4, respectively, that requirements (c) and (d) are satisfied if $\mathfrak{D}$ denotes the class of regular δ-fractions. The proof of Theorem 2.4 turns out to be considerably more complicated than the proof given by Perron [21, p. 111] of the corresponding result for C-fractions. To further illustrate property (d), we offer the following three examples of unique finite regular δ-fraction expansions of rational functions:

$$1+z^2 = (1-z) + \frac{z}{1} - \frac{z}{1} + \frac{z}{1}, \tag{1.3}$$

$$1+z^3 = (1-z) + \frac{z}{(1-z)} + \frac{z}{1} + \frac{z}{1} - \frac{z}{1} + \frac{z}{1} - \frac{z}{1}, \tag{1.4}$$

$$1+z^4 = (1-z) + \frac{z}{(1-z)} + \frac{z}{(1-z)} + \frac{z}{1} - \frac{z}{1} + \frac{z}{1} + \frac{z}{(1-z)} + \frac{z}{1} - \frac{z}{1}. \tag{1.5}$$

In §3 we hope we meet requirement (f) with our δ-fractions by offering seven convergence theorems involving these continued fractions. Using Poincaré's theorem on linear homogeneous difference equations, a version of which we state in Theorem A, we are able to say quite a bit in Theorem 3.1 about the convergence behavior of $\{B_n(z)/B_{n-1}(z)\}$. Here $B_n(z)$ denotes the nth denominator of (1.1). Theorem 3.3 essentially states that the δ-fraction (1.1) with $b_0=1$ converges to an analytic function in a neighborhood of $z=0$, if the sequence $\{d_n\}$ is bounded. It follows from Theorem 3.4 that the δ-fraction (1.1) converges to a meromorphic function in the unit disk $|z|<1$, provided only that $\lim_{n\to\infty} d_n=0$. In §3 we also introduce the concept of a (p,q) *limit periodic δ-fraction*. Theorems 3.5, 3.6 and 3.7 are very useful convergence theorems for limit periodic δ-fractions of types (1,1), (1,2) and (2,1), respectively. We found Theorem 3.5 to be particularly applicable in §4. The various techniques used in the proofs of these convergence theorems can be used to obtain further results of this type. For other recent results involving limit periodic continued fractions, we refer the reader to the work of Thron and Waadeland [27], [28], [29] and to the work of Gill [6], [7].

In §4, we demonstrate that the regular δ-fractions meet requirement (e) by the many examples we give of such continued fraction expansions of classical analytic functions. We use a *method of repeated extensions* along with frequent equivalence transformations to derive many of these expansions. We display several techniques for establishing the validity of these representations. A powerful new technique, based on Poincaré's theorem, allows us in many cases, to establish convergence of a δ-fraction over the whole complex plane (except for poles and cuts) to the analytic function to which it corresponds. Our examples also demonstrate that (p,q) limit periodic behavior is quite common in the expansions of classical functions. It is surprising to this investigator that, in all of the regular δ-fraction expansions given in §4, the δ_n in the partial denominators of these expansions are 0 for $n\geq 2$. We remark in passing that, gathered in the proofs of the theorems of §4, are a large part of the known continued fraction representations of various types for classical functions, analytic at the origin. Also relevant is a paper of Hayden [8] giving other results dealing with continued fraction approximations to functions.

Finally, we assert that the regular δ-fractions meet the remaining requirement (g), not yet considered. However, the only justification for this assertion that we give is to point out the following: The class of regular δ-fractions contains the class of regular C-fractions, and the latter have known connections with the Padé table (see [9, p. 190]).

We refer the reader to Jones and Thron [9, Chapt. 2] for the basic definitions, formulas and properties of continued fractions that are employed in this paper. Other valuable reference books on the subject of continued fractions are those by Perron [21] and Wall [33].

2. Correspondence. We follow the work of Jones and Thron [10], or [9, §5.1] on sequences of meromorphic functions corresponding to a formal Laurent series, in introducing the basic definitions and notation for this section. We call

$$L=c_m z^m+c_{m+1}z^{m+1}+c_{m+2}z^{m+2}+\cdots, \qquad c_m\neq 0, \quad m\geq 0,$$

where the c_m are complex numbers, a *formal power series* (fps). $L=0$ is also called a (fps). We define a function λ on the family of all such power series L as follows:

$$\lambda(L)=\infty \quad \text{if } L=0, \qquad \lambda(L)=m \quad \text{if } L\neq 0.$$

If $f(z)$ is a function analytic at the origin (i.e., analytic in an open disk containing $z=0$), then its Taylor series expansion about $z=0$ will be denoted by $L(f)$. A sequence

$\{R_n(z)\}$ of functions, where each $R_n(z)$ is analytic at the origin, will be said to *correspond to a* (fps) L *at* $z=0$ if

$$\lim_{n\to\infty} \lambda(L-L(R_n))=\infty.$$

If $\{R_n(z)\}$ corresponds to a (fps) L then the *order of correspondence* ν_n of $R_n(z)$ is defined by

$$\nu_n=\lambda(L-L(R_n)).$$

Thus if $\{R_n(z)\}$ corresponds to L, then L and $L(R_n)$ agree term-by-term up to and including the term z^{ν_n-1}. Finally, a continued fraction

$$b_0(z)+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{a_n(z)}{b_n(z)}\right)$$

is said to correspond to a (fps) L if its sequence of approximants corresponds to L.

THEOREM 2.1. *For every formal power series*

$$L_0=1+c_1z+c_2z^2+\cdots$$

there exists a uniquely determined regular δ*-fraction*

$$\mathsf{K}=1-\delta_0z+\frac{d_1z}{1-\delta_1z}+\frac{d_2z}{1-\delta_2z}+\cdots$$

such that K *corresponds to* L_0.

Proof. We define a sequence $\{L_n\}$ of power series each with constant term 1 as follows: If $L_0\equiv 1$, choose $L_1\equiv 1$ and define the δ-fraction K by $\mathsf{K}=1$. If $L_0\not\equiv 1$ and $c_1\neq 0$, choose $\delta_0=0$ and $d_1=c_1$. If $L_0\not\equiv 1$ and $c_1=0$, choose $\delta_1=1$ and $d_1=1$. Then

$$L_0=1-\delta_0z+d_1zL_0^*, \quad \text{where } L_0^*=1+c_1^*z+c_2^*z^2+\cdots,$$

and we define

$$L_1=\frac{1}{L_0^*}=1+c_{1,1}z+c_{1,2}z^2+c_{1,3}z^3+\cdots.$$

If $L_0,L_1,\cdots,L_n$ are defined and have constant term 1, then we define L_{n+1} in the following manner: If $L_n\equiv 1$, choose $L_{n+1}\equiv 1$. If $L_n\not\equiv 1$ and

$$L_n=1+c_{n,1}z+c_{n,2}z^2+\cdots,$$

choose $\delta_n=0$ and $d_{n+1}=c_{n,1}$ if $c_{n,1}\neq 0$. Otherwise, if $c_{n,1}=0$, define

$$\delta_n=1 \quad \text{and} \quad d_{n+1}=1.$$

Then, if $L_n\not\equiv 1$, we have

$$L_n=1-\delta_nz+d_{n+1}zL_n^*, \quad \text{where } L_n^*=1+c_{n,1}^*z+c_{n,2}^*z^2+\cdots.$$

In this case, we define

$$L_{n+1}=\frac{1}{L_n^*}=1+c_{n+1,1}z+c_{n+1,2}z^2+\cdots.$$

Hence, it follows by induction that L_n is defined for all $n\geq 0$. It is easy to see that if $L_n\equiv 1$ for some n then $L_k\equiv 1$ for all $k\geq n$. If this should happen, let m be the least

value of n such that $L_n \equiv 1$. Then we have

$$L_k = L(1-\delta_k z) + \frac{L(d_{k+1}z)}{L_{k+1}}, \qquad 0 \le k \le m-1,$$

where $d_{k+1} = 1$ if $\delta_k = 1$.

If we set $a_k(z) = d_k z \neq 0$ and $b_k(z) = 1-\delta_k z$, then it follows from [10, Thm. 2] that

$$L_0 = L\left(1-\delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \cdots + \frac{d_{m-1}z}{1-\delta_{m-1}z} + \frac{d_m z}{1}\right),$$

and K corresponds to L_0 if

$$\mathsf{K} = 1-\delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \cdots + \frac{d_{m-1}z}{1-\delta_{m-1}z} + \frac{d_m z}{1}.$$

In the remaining case, where $L_n \not\equiv 1$ for all n, we have shown that there exist sequences $\{\delta_k\}$ and $\{d_k\}$, where $d_k \neq 0$, $\delta_k = 0$ or 1, and $d_{k+1} = 1$ if $\delta_k = 1$ such that

$$L_k = L(1-\delta_k z) + \frac{L(d_{k+1}z)}{L_{k+1}}, \qquad k \ge 0.$$

Again, if we set $a_k(z) = d_k z$ and $b_k(z) = 1-\delta_k z$, it follows from [9, Thm. 2] that K corresponds to L_0, where

$$\mathsf{K} = 1-\delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \frac{d_2 z}{1-\delta_2 z} + \cdots.$$

With this our proof is complete.

THEOREM 2.2. *Each finite δ-fraction*

$$\mathsf{K}_n = 1-\delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \cdots + \frac{d_{n-1}z}{1-\delta_{n-1}z} + \frac{d_n z}{1}$$

and each infinite δ-fraction

$$\mathsf{K}_\infty = 1-\delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \frac{d_2 z}{1-\delta_2 z} + \cdots$$

corresponds to a uniquely determined power series

$$L_0 = 1 + c_1 z + c_2 z^2 + \cdots$$

The order of correspondence ν_k *of* K_∞ *is* $k+1$, *and* ν_k *of* K_n *is* $k+1$ *if* $0 \le k < n$ *and* ∞ *if* $k \ge n$.

Proof. Let the sequences $\{R_n(z)\}$ and $\{S_n(z)\}$ of functions, analytic at the origin, be defined by

$$R_n(z) = \frac{A_n(z)}{B_n(z)}, \quad n \ge 0, \qquad S_k(z) = \frac{A_k^*(z)}{B_k^*(z)}, \quad 0 \le k \le n, \qquad S_k(z) \equiv \frac{A_n^*(z)}{B_n^*(z)}, \quad k > n,$$

where $A_n(z)/B_n(z)$ is the nth approximant of K_∞ and $A_k^*(z)/B_k^*(z)$ is the kth approximant of K_n. Since

$$R_{n+1}(z) - R_n(z) = \frac{(-1)^n d_1 d_2 \cdots d_{n+1} z^{n+1}}{B_n(z) B_{n+1}(z)}$$

and $B_n(0) \equiv 1$, it follows that

$$\nu_n = \lambda(L(R_{n+1}) - L(R_n)) = n+1 \to \infty$$

monotonically as $n \to \infty$. Also

$$S_{k+1}(z) - S_k(k) = \begin{cases} \dfrac{(-1)^k d_1 d_2 \cdots d_{k+1} z^{k+1}}{B_k^*(z) B_{k+1}^*(z)} & \text{if } k < n, \\ 0 & \text{if } k \geq n. \end{cases}$$

Hence, in this case,

$$\nu_k = \lambda(L(S_{k+1}) - L(S_k)) = \begin{cases} k+1 & \text{if } k < n, \\ \infty & \text{if } k \geq n, \end{cases}$$

and again $\lim_{k\to\infty} \nu_k = \infty$. Thus we have met the requirements of [10, Thm. 1] and our theorem follows immediately from this result.

THEOREM 2.3. (A) *Two regular infinite δ-fractions*

$$\mathsf{K} = 1 - \delta_0 z + \mathop{\mathsf{K}}_{k=1}^{\infty} \left(\frac{d_k}{1 - \delta_k z} \right) \quad \textit{and} \quad \mathsf{K}^* = 1 - \delta_0^* z + \mathop{\mathsf{K}}_{k=1}^{\infty} \left(\frac{d_k^* z}{1 - \delta_k^* z} \right)$$

correspond to the same power series

$$L_0 = 1 + c_1 z + c_2 z^2 + \cdots$$

if and only if

$$\delta_k = \delta_k^*, \quad k \geq 0, \qquad d_k = d_k^*, \quad k \geq 1.$$

(B) *A regular infinite δ-fraction* K^* *and a regular finite δ-fraction*

$$\mathsf{K}_n = 1 - \delta_0 z + \frac{d_1 z}{1 - \delta_1 z} + \cdots + \frac{d_{n-1} z}{1 - \delta_{n-1} z} + \frac{d_n z}{1}$$

correspond to the same power series L_0 *if and only if*

$$\delta_k = \delta_k^*, \quad 0 \leq k \leq n-1, \qquad \delta_k^* = 1, \quad k \geq n,$$
$$d_k = d_K^*, \quad 1 \leq k \leq n, \qquad d_k^* = 1, \quad k \geq n+1.$$

(C) *A regular finite δ-fraction* K_n *and a regular finite δ-fraction*

$$\mathsf{K}_m^* = 1 - \delta_0^* z + \frac{d_1^* z}{1 - \delta_1^* z} + \cdots + \frac{d_{m-1}^* z}{1 - \delta_{m-1}^* z} + \frac{d_m^* z}{1}$$

correspond to the same power series L_0 *if and only if*

$$n = m, \qquad \delta_k = \delta_k^*, \quad 0 \leq k \leq n-1, \qquad d_k = d_k^*, \quad 1 \leq k \leq n.$$

Proof. We prove part (A) first. Let $A_n(z)/B_n(z)$ and $A_n^*/B_n^*(z)$ for all $n \geq 0$ denote the nth approximants of K and K^*, respectively. Suppose K and K^* correspond to the same power series L_0. Then $\delta_0 = \delta_0^*$, for if not, consider

$$\frac{A_1}{B_1} - \frac{A_1^*}{B_1^*} = \frac{d_1 z}{1 - \delta_1 z} + (\delta_0^* - \delta_0) z - \frac{d_1^* z}{1 - \delta_1^* z} = (d_1 + \delta_0^* - \delta_0 - d_1^*) z + \cdots.$$

By Theorem 2.2, $\lambda(L_0 - L(A_1/B_1)) = \lambda(L_0 - L(A_1^*/B_1^*)) = 2$. Hence

$$d_1 + \delta_0^* - \delta_0 - d_1^* = 0.$$

Since $\delta_0^* - \delta_0 \neq 0$, either $\delta_0^* = 1$ and $\delta_0 = 0$ or $\delta_0 = 1$ and $\delta_0^* = 0$. But then either $d_1 = 0$ or $d_1^* = 0$; neither of which can happen. If K and K^* still are not identical, let ν be the smallest value of n for which one of the conditions

$$d_n \neq d_n^*, \qquad \delta_n \neq \delta_n^*$$

is satisfied. From the fundamental formulas we obtain

$$\frac{A_\nu}{B_\nu}-\frac{A_{\nu-1}}{B_{\nu-1}}=\frac{(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}d_\nu z^\nu}{B_\nu B_{\nu-1}}$$

and

$$\frac{A_\nu^*}{B_\nu^*}-\frac{A_{\nu-1}^*}{B_{\nu-1}^*}=\frac{(-1)^{\nu-1}d_1^*d_2^*\cdots d_{\nu-1}^*d_\nu^* z^\nu}{B_\nu^* B_{\nu-1}^*}.$$

Since $d_k=d_k^*$, $\delta_k=\delta_k^*$, $A_k=A_k^*$, $B_k=B_k^*$, $k\leq\nu-1$, it follows that

$$\begin{aligned}\frac{A_\nu}{B_\nu}-\frac{A_\nu^*}{B_\nu^*}&=\frac{(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}z^\nu}{B_{\nu-1}}\left(\frac{d_\nu}{B_\nu}-\frac{d_\nu}{B_\nu^*}\right)\\&=(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}(d_\nu-d_\nu^*)z^\nu+\cdots.\end{aligned}$$

From Theorem 2.2 we have that the order of correspondence is $\nu+1$ for both A_ν/B_ν and A_ν^*/B_ν^*. Therefore it follows from the last equation that

$$d_\nu=d_\nu^*$$

and

$$\begin{aligned}\frac{A_\nu}{B_\nu}-\frac{A_\nu^*}{B_\nu^*}&=\frac{(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}d_\nu z^\nu}{B_{\nu-1}}\left(\frac{1}{B_\nu}-\frac{1}{B_\nu^*}\right)\\&=\frac{(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}d_\nu z^{\nu+1}}{B_\nu B_\nu^*}(\delta_\nu-\delta_\nu^*)\\&=(-1)^{\nu-1}d_1d_2\cdots d_{\nu-1}d_\nu(\delta_\nu-\delta_\nu^*)z^{\nu+1}+\cdots.\end{aligned}$$

With the aid of this result we obtain

$$\begin{aligned}\frac{A_{\nu+1}}{B_{\nu+1}}-\frac{A_{\nu+1}^*}{B_{\nu+1}^*}&=\left(\frac{A_{\nu+1}}{B_{\nu+1}}-\frac{A_\nu}{B_\nu}\right)+\left(\frac{A_\nu}{B_\nu}-\frac{A_\nu^*}{B_\nu^*}\right)+\left(\frac{A_\nu^*}{B_\nu^*}-\frac{A_{\nu+1}^*}{B_{\nu+1}^*}\right)\\&=\frac{(-1)^\nu d_1\cdots d_{\nu+1}z^{\nu+1}}{B_{\nu+1}B_\nu}+\left(\frac{A_\nu}{B_\nu}-\frac{A_\nu^*}{B_\nu^*}\right)+\frac{(-1)^{\nu+1}d_1\cdots d_\nu d_{\nu+1}^* z^{\nu+1}}{B_{\nu+1}^*B_\nu^*}\\&=(-1)^\nu(d_1\cdots d_\nu)(d_{\nu+1}-\delta_\nu+\delta_\nu^*-d_{\nu+1}^*)z^{\nu+1}+\cdots.\end{aligned}$$

Again, since $L(A_{\nu+1}/B_{\nu+1})$ and $L(A_{\nu+1}^*/B_{\nu+1}^*)$ agree in powers of z up through $z^{\nu+1}$, it follows that

$$d_{\nu+1}-\delta_\nu+\delta_\nu^*-d_{\nu+1}^*=0.$$

By an argument similar to the one used to show $\delta_0=\delta_0^*$, it follows that $\delta_\nu=\delta_\nu^*$ and $d_{\nu+1}=d_{\nu+1}^*$. Thus we have shown that $d_\nu=d_\nu^*$ and $\delta_\nu=\delta_\nu^*$, contradicting our assumption that at least one of these equations fails to hold. Hence, if K and K^* correspond to the same L_0, they must be identical and our proof of part (A) is complete.

To prove part (B) it is sufficient to prove that K^* corresponds to $L_0=L(\mathsf{K}_n)$, for then by part (A) any other infinite δ-fraction corresponding to L_0 is identical to K^*. Let A_k/B_k, $0\leq k\leq n$, denote the kth approximant of K_n. Then $A_k/B_k=A_k^*/B_K^*$,

$0 \le k \le n-1$. With the aid of the fundamental formulas [21, p. 1], it is easily verified that

$$\frac{A^*_{n+m}}{B^*_{n+m}} = \frac{A_n + (P_m/Q_m)A_{n-1}}{B_n + (P_m/Q_m)B_{n-1}}, \qquad m \ge 0,$$

where $A_n/B_n = \mathsf{K}_n$ and P_m/Q_m is the mth approximant of

$$-z + \frac{z}{1-z} + \frac{z}{1-z} + \frac{z}{1-z} + \cdots.$$

Since $P_0 = -z$, $P_1 = z^2$ and $P_m = P_{m-1}(1-z) + zP_{m-2}$ for $m \ge 2$, it follows by induction that

$$P_m(z) = (-1)^{m-1} z^{m+1}, \qquad m \ge 0.$$

It is also easy to see that $Q_m(0) = 1$ for all $m \ge 0$. Now

$$\begin{aligned}
\frac{A^*_{n+m}}{B^*_{n+m}} - \frac{A_n}{B_n} &= \frac{A_n + (P_m/Q_m)A_{n-1}}{(P_m/Q_m)B_{n-1}} - \frac{A_n}{B_n} \\
&= \frac{(P_m/Q_m)(A_{n-1}B_n - A_nB_{n-1})}{B_n + (P_m/Q_m)B_{n-1}} \\
&= \frac{P_m(-1)^n d_1 \cdots d_n z^n}{Q_m(B_n + (P_m/Q_m)B_{n-1})} \\
&= (-1)^{n+m+1} d_1 \cdots d_n z^{n+m+1} + \cdots.
\end{aligned}$$

Hence

$$\lambda[L(A^*_{n+m}/B^*_{n+m}) - L_0] = n+m+1 \to \infty \quad \text{as } m \to \infty.$$

This, coupled with the fact that

$$\lambda\left(L\left(\frac{A^*_k}{B^*_k}\right) - L\left(\frac{A_k}{B_k}\right)\right) = \infty, \qquad 0 \le k \le n-1,$$

guarantees that K^* corresponds to L_0.

With the aid of parts (A) and (B), we are now in a position to prove part (C). By part (B)

$$\mathsf{K}^* = 1 - \delta^*_0 z + \frac{d^*_1 z}{1-\delta^*_1 z} + \cdots + \frac{d^*_{m-1} z}{1-\delta^*_{m-1} z} + \frac{d^*_m z}{1-z} + \frac{z}{1-z} + \frac{z}{1-z} + \cdots$$

corresponds to $L(\mathsf{K}^*_m)$ and

$$\mathsf{K} = 1 - \delta_0 z + \frac{d_1 z}{1-\delta_1 z} + \cdots + \frac{d_{n-1} z}{1-\delta_{n-1} z} + \frac{d_n z}{1-z} + \frac{z}{1-z} + \frac{z}{1-z} + \cdots$$

corresponds to $L(\mathsf{K}_n)$. Since we have assumed that $L(\mathsf{K}^*_m) = L(\mathsf{K}_n) = L_0$, it follows from part (A) that K is identical to K^* and therefore our assertion in part (C) is true.

THEOREM 2.4. *A power series*

$$L_0 = 1 + c_1 z + c_2 z^2 + \cdots$$

is the Taylor series about the origin of a rational function

$$R(z) = \frac{1 + a_1 z + \cdots + a_n z^n}{1 + b_1 z + \cdots + b_m z^m}$$

if and only if there exists a finite regular δ-fraction

$$\mathsf{K}=1-\delta_0 z+\frac{d_1 z}{1-\delta_1 z+\cdots+\dfrac{d_{n-1}z}{1-\delta_{n-1}z+\dfrac{d_n z}{1}}}$$

such that K *corresponds to* L_0.

Proof. Let the degree of any polynomial $P(z)$ be denoted by $\hat{P}$. We shall prove the above theorem by mathematical induction, where our nth induction statement is that all rational functions of the form $P(z)/Q(z)$, where $P(0)=Q(0)=1$ and $\max(\hat{P},\hat{Q})=n$, have a δ-fraction expansion of the type asserted in the theorem. If $n=0$, then $P(z)/Q(z)\equiv 1$ and we have only to choose $\mathsf{K}=1$. If $n=1$, then we have only to consider the three types of rational functions $R_1(z)=1+az$, $R_2(z)=1/(1+bz)$, or $R_3(z)=(1+az)/(1+bz)$, where $a\neq 0$, $b\neq 0$. The desired δ-fraction expansions for R_1 and R_2 are clearly

$$1+\frac{az}{1} \quad \text{and} \quad 1-\frac{bz}{1+\dfrac{bz}{1}},$$

respectively. If $a=b$ in R_3, choose $\mathsf{K}=1$; otherwise, the desired expansion for R_3 is easily seen to be

$$1+\frac{(a-b)z}{1+\dfrac{bz}{1}}.$$

Now assume that our theorem is true for all rational functions $R(z)=P(z)/Q(z)$ satisfying $P(0)=Q(0)=1$ and $\max(\hat{P},\hat{Q})=k$ for some k, $0\leq k\leq n$. Let $R_0(z)=P_0(z)/Q_0(z)$ be an arbitrary rational function satisfying $P_0(0)=Q_0(0)=1$ and $\max(\hat{P}_0,\hat{Q}_0)=n+1$. To complete our proof we shall show that either $R_0(z)\equiv 1$, in which case we choose $\mathsf{K}=1$, or $R_0(z)\not\equiv 1$ can be expressed in one of the following six forms in a neighborhood of $z=0$:

(a) $$R_0(z)=1+\frac{d_1 z}{R_1(z)},$$

(b) $$R_0(z)=1+\frac{d_1 z}{1+\dfrac{d_2 z}{R_1(z)}},$$

(c) $$R_0(z)=(1-z)+\underbrace{\frac{z}{1-z+\cdots+\dfrac{z}{1-z}}}_{\alpha_1\text{-2 terms}}+\frac{z}{1}+\frac{d_1 z}{R_1(z)},$$

(d) $$R_0(z)=1+\frac{d_1 z}{1-z+\underbrace{\dfrac{z}{1-z+\cdots+\dfrac{z}{1-z}}}_{\alpha_2\text{-2 terms}}+\dfrac{z}{1}+\dfrac{d_2 z}{R_1(z)}},$$

(e) $$R_0(z)=(1-z)+\underbrace{\frac{z}{1-z+\cdots+\dfrac{z}{1-z}}}_{\alpha_1\text{-2 terms}}+\frac{z}{1}+\frac{d_1 z}{1}+\frac{d_2 z}{R_1(z)},$$

(f) $$R_0(z)=(1-z)+\underbrace{\frac{z}{1-z+\cdots+\dfrac{z}{1-z}}}_{\alpha_1\text{-2 terms}}+\frac{z}{1}+\frac{d_1 z}{1-z}+\underbrace{\frac{z}{1-z+\cdots+\dfrac{z}{1-z}}}_{\alpha_2\text{-2 terms}}+\frac{z}{1}+\frac{d_2 z}{R_1(z)},$$

where $d_1\neq 0$, $d_2\neq 0$ are complex numbers, $\alpha_1\geq 2$, $\alpha_2\geq 2$ are positive integers, and $R_1(z)=P_1(z)/Q_1(z)$, where $P_1(0)=Q_1(0)=1$ and $\max(\hat{P}_1,\hat{Q}_1)\leq n$.

We introduce a formula that we shall need in the course of our proof. If

$$F(z)=1+\frac{az^{\alpha}}{H(z)}, \tag{2.1}$$

where $a\neq 0$, α is an integer ≥ 2, and $H(z)$ is a rational function satisfying $H(0)=1$, then in a neighborhood of $z=0$ it is easy to see that

$$F(z)=(1-z)+\frac{z}{1-\dfrac{az^{\alpha-1}}{az^{\alpha-1}+H(z)}}. \tag{2.2}$$

Now let

$$R_0(z)=\frac{P_0(z)}{Q_0(z)},\quad P_0(0)=Q_0(0)=1,\quad \max(\hat{P}_0,\hat{Q}_0)=n+1.$$

If $R_0\not\equiv 1$, then

$$R_0(z)=1+\frac{a_1z^{\alpha_1}}{R_1(z)},$$

where $a_1\neq 0$, α_1 is a positive integer and $R_1(z)=Q_0(z)/Q_1(z)$, $Q_1(0)=1$ and $a_1z^{\alpha_1}Q_1(z)=P_0(z)-Q_0(z)$. We note that

$$\hat{Q}_1\leq\max(\hat{P}_0-\alpha_1,\hat{Q}_0-\alpha_1)\leq\max(\hat{P}_0-1,\hat{Q}_0-1)\leq n.$$

Suppose $\alpha_1=1$. Then if $\max(\hat{Q}_0,\hat{Q}_1)\leq n$, $R_0(z)$ is in the form (a). Otherwise, suppose $\max(\hat{Q}_1,\hat{Q}_1)=n+1$; then

$$R_0(z)=1+\frac{a_1z}{1+\dfrac{a_2z^{\alpha_2}}{R_2(z)}},$$

where $a_2\neq 0$, α_2 is a positive integer and $R_2(z)=Q_1(z)/P_1(z)$, where $a_2z^{\alpha_2}P_1(z)=Q_0(z)-Q_1(z)$ and $P_1(0)=1$. We have

$$\hat{P}_1\leq\max(\hat{Q}_0-\alpha_2,\hat{Q}_1-\alpha_2)\leq\max(\hat{Q}_0-\alpha_2,\hat{P}_0-\alpha_1-\alpha_2)\leq n.$$

Thus $R_0(z)$ is of the form (b) if $\alpha_2=1$. If $\alpha_2>1$, then by repeated use of (2.1) and (2.2) we can write

$$R_0(z)=1+\frac{a_1z}{1-z}+\underbrace{\frac{z}{1-z}+\cdots+\frac{z}{1-z}}_{\alpha_2\text{-}2\text{ terms}}+\frac{z}{1}+\frac{(-1)^{\alpha_2-1}a_2z}{H_2(z)+R_2(z)},$$

where

$$H_2(z)=\sum_{k=1}^{\alpha_2-1}(-1)^{k+1}z^{\alpha_2-k}.$$

We write

$$H_2(z)+R_2(z)=R_3(z),$$

where $R_3(z)=P_2(z)/P_1(z)$ with $P_2(z)=Q_1(z)+P_1(z)H_2(z)$. We note that $P_2(0)=1$ and

$$\hat{P}_2\leq\max(\hat{Q}_1,\hat{P}_1+\alpha_2-1)\leq\max(\hat{Q}_0,\hat{P}_0)-1\leq n.$$

Hence in this case $R_0(z)$ is of the form (d).

Now suppose $\alpha_1>1$. Then by repeated application of (2.1) and (2.2) we obtain

$$R_0(z)=1-z+\underbrace{\frac{z}{1-z}+\cdots+\frac{z}{1-z}}_{\alpha_1\text{-}2\text{ terms}}+\frac{z}{1}+\frac{(-1)^{\alpha_1-1}a_1z}{H_1(z)+R_1(z)},$$

where

$$H_1(z)=\sum_{k=1}^{\alpha_1-1}(-1)^{k+1}z^{\alpha_1-k}.$$

We set $H_1(z)+R_1(z)=R_4(z)=Q_2(z)/Q_1(z)$, where $Q_2(z)=Q_0(z)+Q_1(z)H_1(z)$. Also, $Q_2(0)=1$ and $\hat{Q}_2\le\max(\hat{Q}_0,\hat{Q}_1+\alpha_1-1)\le\max(\hat{Q}_0,\hat{P}_0-1)$. If $\max(\hat{Q}_0,\hat{Q}_1)\le n$, then $R_0(z)$ is of the form (c). Otherwise,

$$R_0(z)=1-z+\underbrace{\frac{z}{1-z}+\cdots+\frac{z}{1-z}}_{\alpha_1\text{-2 terms}}+\frac{z}{1}+\frac{(-1)^{\alpha_1-1}a_1z}{1}+\frac{a_3z^{\alpha_3}}{R_5(z)},$$

where $R_5(z)=Q_1(z)/P_3(z)$, $a_3z^{\alpha_3}P_3(z)=Q_2(z)-Q_1(z)$, $P_3(0)=1$, $a_3\neq0$, α_3 is a positive integer and

$$\hat{P}_3\le\max(\hat{Q}_2-\alpha_3,\hat{Q}_1-\alpha_3)\le\max(\hat{Q}_0-\alpha_3,\hat{P}_0-1-\alpha_3)\le n.$$

If $\alpha_3=1$, $R_0(z)$ is of the form (e). If $\alpha_3>1$, then

$$R_0(z)=1-z+\underbrace{\frac{z}{1-z}+\cdots+\frac{z}{1-z}}_{\alpha_1\text{-2 terms}}+\frac{z}{1}+\frac{(-1)^{\alpha_1-1}a_1z}{1-z}$$

$$+\underbrace{\frac{z}{1-z}+\cdots+\frac{z}{1-z}}_{\alpha_3\text{-2 terms}}+\frac{z}{1}+\frac{(-1)^{\alpha_3-1}a_3z}{H_3(z)+R_5(z)},$$

where

$$H_3(z)=\sum_{k=1}^{\alpha_3-1}(-1)^{k+1}z^{\alpha_3-k}.$$

Now $H_3(z)=P_3(z)/Q_3(z)$ with $a_3z^{\alpha_3}Q_3(z)=Q_1(z)-P_3(z)$, $Q_3(0)=1$, and

$$\hat{Q}_3\le\max(\hat{Q}_1-\alpha_3,\hat{P}_3-\alpha_3)\le\max(\hat{Q}_0,\hat{P}_0)-1\le n.$$

Hence $R_0(z)$ is of the form (f). Since we have now covered all cases our proof is complete.

THEOREM 2.5. *A regular infinite δ-fraction*

$$\mathsf{K}=1-\delta_0z+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{d_nz}{1-\delta_nz}\right)$$

corresponds to the Taylor series expansion about $z=0$ of a rational function

$$R(z)=\frac{1+a_1z+\cdots+a_nz^n}{1+b_1z+\cdots+b_mz^m}$$

if and only if there exists an integer $N\ge0$ such that $\delta_n=1$ if $n\ge N$ and $d_n=1$ if $n\ge N+1$.

Proof. By Theorem 2.4 there exists a finite δ-fraction

$$\mathsf{K}_n=1-\delta_0z+\frac{d_1z}{1-\delta_1z}+\cdots+\frac{d_{n-1}z}{1-\delta_{n-1}z}+\frac{d_nz}{1}$$

such that K_n corresponds to $L(R(z))$. Then by part (B) of Theorem 2.3, the infinite δ-fraction

$$\mathsf{K}=1-\delta_0 z+\frac{d_1 z}{1-\delta_1 z}+\cdots+\frac{d_{n-1}z}{1-\delta_{n-1}z}+\frac{d_n z}{1-z}+\frac{z}{1-z}+\frac{z}{1-z}+\cdots$$

also corresponds to $L(R(z))$. By part (A) of the Theorem 2.3 any other infinite δ-fraction corresponding to $L(R(z))$ must be identical to K, so we have our desired result.

3. Convergence. We shall need the following theorem, a version of which was first given by Poincaré [25] in 1885.

THEOREM A (Poincaré). *For $n=1,2,3,\cdots$, let D_n be a nontrivial solution of the homogeneous linear difference equation*

$$D_n=b_n D_{n-1}+a_n D_{n-2}, \tag{3.1}$$

where

$$\lim_{n\to\infty} a_n=a, \qquad \lim_{n\to\infty} b_n=b.$$

Let the roots x_1 and x_2 of the characteristic equation

$$x^2-bx-a=0 \tag{3.2}$$

satisfy

$$|x_1|>|x_2|. \tag{3.3}$$

Then

$$\lim_{n\to\infty}\frac{D_n}{D_{n-1}}=x_k$$

where x_k is one of the roots of (3.2).

To fix the ideas of his method of proof Poincaré sketched the proof of a similar result for third order linear difference equations. Perron [22], [23], [24] gave extensions of Poincaré's theorem and proofs for the general nth order case through three separate papers, the first two of which appeared in 1909 and the last in 1921. Treatments in English of the theorems of Poincaré and Perron on finite differences may be found in the books of Gel'fond [5] and Milne-Thomson [19] on the calculus of finite differences.

An immediate application of the Poincaré theorem to the theory of δ-fractions is the following result:

THEOREM 3.1. *Let*

$$b_0-\delta_0 z+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{d_n z}{1-\delta_n z}\right) \tag{3.4}$$

be a δ-fraction satisfying

$$\lim_{n\to\infty} d_n=d, \qquad \lim_{n\to\infty}\delta_n=0. \tag{3.5}$$

Let

$$R_d=\left\{\frac{-t}{4d}: t\geq 1\right\} \quad \text{if } d\neq 0, \qquad R_d=\mathbb{C} \quad \text{if } d=0.$$

For $z\in R_d$ let $\sqrt{zd+1/4}$ denote the square root with positive real part.

(A) *If* $z \in R_d$ *and if* $B_n(z)$ *denotes the nth denominator of* (3.4), *then*

$$\lim_{n\to\infty} \frac{B_n(z)}{B_{n-1}(z)} = x(z),$$

where $x(z)$ *is one of the two values*

$$\frac{1}{2} \pm \sqrt{zd+\frac{1}{4}}\,.$$

(B) *If condition* (3.5) *is replaced by*

$$\lim_{n\to\infty} d_n = d \neq 0, \qquad |d_n - d| \leq \frac{\theta}{8M}, \quad \delta_n \equiv 0, \quad n \geq 1, \tag{3.6}$$

and if $z \in H(\theta, M, d)$, *where* $0<\theta<1$, $M>0$ *and*

$$H(\theta,M,d) = \left\{z: |z| - \left|z+\frac{1}{4d}\right| \leq \frac{1-\theta}{4|d|}\right\} \cap \{z: |z| \leq M\},$$

then

$$\lim_{n\to\infty} \frac{B_n(z)}{B_{n-1}(z)} = \frac{1}{2} + \sqrt{zd+\frac{1}{4}}\,.$$

(C) *Finally, if*

$$\lim_{n\to\infty} d_n = 0, \qquad |d_n| \leq \frac{1}{4M}, \quad \delta_n \equiv 0, \quad n \geq 1, \tag{3.7}$$

and if $|z| \leq M$, *then*

$$\lim_{n\to\infty} \frac{B_n(z)}{B_{n-1}(z)} = 1.$$

Proof. The nth denominator B_n of (3.4) satisfies the recurrence relation

$$B_n = (1-\delta_n z)B_{n-1} + zd_n B_{n-2}.$$

Since $\lim_{n\to\infty}(1-\delta_n z) = 1$ and $\lim_{n\to\infty} d_n z = zd$, it follows from Poincaré's theorem that $\{h_n(z)\}$, where $h_n(z) = B_n(z)/B_{n-1}(z)$, converges to one of the two roots of the quadratic equation

$$x^2 - x - zd = 0, \tag{3.8}$$

provided these roots have unequal moduli. It is easily seen that this is the case if and only if $z \in R_d$. Hence (A) is true.

To aid in verifying (B) we set the roots of (3.8) equal to $x_1(z)$ and $x_2(z)$ where

$$x_1(z) = \frac{1}{2} + \sqrt{zd+\frac{1}{4}} \quad \text{and} \quad x_2(z) = \frac{1}{2} - \sqrt{zd+\frac{1}{4}}\,.$$

We have that $|x_1(z)| > |x_2(z)|$ for all $z \in R_d$ and therefore for all $z \in H(\theta, M, d)$, since $H(\theta, M, d) \subset R_d$. The latter inclusion is easily verified after observing that the boundary of $H(\theta, M, d)$ consists of an arc of a circle and an arc of a hyperbola whose axis contains the cut omitted from R_d. Thus it follows from part (A) that $h_n(z) \to x_1(z)$ or $h_n(z) \to x_2(z)$ as $n \to \infty$. Under the additional assumptions in (B), we shall show that

$\lim_{n\to\infty} h_n(z)=x_1(z)$ by showing that $\{h_n(z)\}$ cannot converge to $x_2(z)$. In fact we shall show by induction on n that $|h_n(z)-x_2(z)|\geq D(z)/2$, where

$$D(z)=|x_1(z)|-|x_2(z)|>0.$$

Since, $B_0(z)=B_1(z)=1$, we have

$$|h_1(z)-x_2(z)|=|1-x_2(z)|=|x_1(z)|>\frac{D(z)}{2}.$$

Furthermore,

$$\frac{D^2(z)}{4}=\frac{1}{2}\left[\left|\frac{1}{4}+zd\right|+\frac{1}{4}-|z|d\right],$$

and it follows from the assumptions of part (B) that

$$|d_n-d||z|\leq\frac{D^2(z)}{4} \quad \text{for all } n\geq 1.$$

Now assume $|h_n(z)-x_2(z)|\geq D(z)/2$ for some n. Then

$$\begin{aligned}
|h_{n+1}(z)-x_2(z)| &= \left|1-x_2(z)+\frac{d_{n+1}z}{h_n(z)}\right| \\
&= \left|x_1(z)+\frac{d_{n+1}z}{h_n(z)}\right| \\
&= \frac{|x_1(z)(h_n(z)-x_2(z))+(d_{n+1}-d)z|}{|h_n(z)|} \\
&\geq \frac{|x_1(z)||h_n(z)-x_2(z)|-|d_{n+1}-d||z|}{|h_n(z)-x_2(z)|+|x_2(z)|} \\
&\geq \frac{|x_1(z)|D(z)/2-D^2(z)/4}{D(z)/2+|x_2(z)|}=\frac{D(z)}{2},
\end{aligned}$$

and our induction argument is complete.

Part (C) is proved in a similar manner. We first note that $x_1(z)=1$ and $x_2(z)=0$ in this case. Via Poincaré (Theorem A) it is true that $h_n(z)\to 0$ or $h_n(z)\to 1$ as $n\to\infty$. We shall show that $|h_n(z)|\geq\frac{1}{2}$ for all $n\geq 1$, so that $h_n\to 1$. Here

$$|h_1(z)|=1>\frac{1}{2}$$

and, assuming $|h_n(z)|\geq\frac{1}{2}$,

$$|h_{n+1}(z)|=\left|1+\frac{d_{n+1}z}{h_n(z)}\right|\geq 1-\frac{|d_{n+1}||z|}{|h_n(z)|}\geq 1-2|d_{n+1}||z|\geq 1-2\left(\frac{1}{4M}\right)M=\frac{1}{2}.$$

This completes the proof of the theorem.

The following theorem is an adaptation to δ-fractions of some convergence results given by Jones and Thron [10] dealing with sequences of meromorphic functions.

THEOREM 3.2. *Let the infinite δ-fraction*

$$\mathsf{K}=1-\delta_0 z+\frac{d_1 z}{1-\delta_1 z}+\frac{d_2 z}{1-\delta_2 z}+\cdots$$

correspond to the power series

$$L_0 = 1 + c_1 z + c_2 z^2 + \cdots .$$

Let D be a domain containing a neighborhood of the origin. Then:

(A) K *converges uniformly on every compact subset of D if and only if its sequence of approximants* $\{A_n(z)/B_n(z)\}$ *is uniformly bounded on every compact subset of D.*

(B) *If* K *converges uniformly on every compact subset of D, then* $f(z) = \lim_{n\to\infty} A_n(z)/B_n(z)$ *is analytic in D and* L_0 *is the Taylor series expansion of* $f(z)$ *about* $z=0$.

(C) *Two infinite subsequences of approximants of* K *which converge uniformly on every compact subset of D converge to the same analytic function in D.*

Proof. Parts (A) and (B) follow immediately from [10, Thm 4′]. Part (C) may be established as follows: Let $\{f_{n_k}(z)\}$ and $\{f_{m_k}(z)\}$ be two subsequences of approximants of K which converge uniformly on every compact subset of D. Since each of these subsequences corresponds to L_0, it follows from [10, Thm. 4′] that $\lim_{k\to\infty} f_{n_k}(z) = F_n(z)$ and $\lim_{k\to\infty} f_{m_k}(z) = F_m(z)$, where $F_n(z)$ and $F_m(z)$ are analytic in D and have the property that L_0 is the Taylor expansion of each about the origin. Since D is a domain, it follows from the identity theorem for analytic functions that $F_n(z) = F_m(z)$ for all z in D.

The next theorem shows that a δ-fraction converges uniformly in a neighborhood of the origin to an analytic function if $\{\delta_n\}$ is an arbitrary sequence of zeros and ones, provided only that the sequence $\{d_n\}$ is bounded.

THEOREM 3.3. *If the coefficients* d_n *of the infinite δ-fraction*

$$\mathsf{K} = 1 - \delta_0 z + \mathop{\mathsf{K}}_{n=1}^{\infty} \left(\frac{d_n z}{1 - \delta_n z} \right)$$

satisfy the inequality

$$0 < |d_n| \le M,$$

then K *converges uniformly in the disk* $|z| \le (\sqrt{1+M} + \sqrt{M})^{-2}$ *to a function* $f(z)$ *which is analytic in the interior of this disk.*

Proof. With the aid of an equivalence transformation, K can be written in the form

$$\mathsf{K} = 1 - \delta_0 z + \mathop{\mathsf{K}}_{n=1}^{\infty} \left(\frac{d_n E_n(z)}{1} \right),$$

where for each n, $E_n(z)$ is one of the three functions

$$z, \quad \frac{z}{1-z}, \quad \frac{z}{(1-z)^2},$$

provided $z \neq 1$. If $|z| \le r$, $0 < r < 1$, then

$$|z| \le \frac{|z|}{1-|z|} \le \frac{|z|}{(1-|z|)^2} \le \frac{r}{(1-r)^2}$$

from which we also derive

$$\frac{|z|}{|1-z|} \le \frac{r}{(1-r)^2} \quad \text{and} \quad \frac{|z|}{|1-z|^2} \le \frac{r}{(1-r)^2},$$

since $|1-z| \ge 1-|z|$.

We impose the further restriction on r that it satisfy the equation

$$\frac{r}{(1-r)^2}=\frac{1}{4M}.$$

We solve this equation for the value of r in the interval $0<r<1$ to obtain

$$r=\left(\sqrt{1+M}+\sqrt{M}\right)^{-2}.$$

By the convergence neighborhood theorem in Jones and Thron [9, p. 108], Perron [21, Satz. 2.25, p. 64] with $p_2\equiv 2$ or by Wall [33, Thm. 10.1, p. 42], K converges uniformly if $|d_nE_n(z)|\leq\frac{1}{4}$. Since the latter inequality is satisfied whenever $|d_n|\leq M$ and $|z|\leq(\sqrt{1+M}+\sqrt{M})^{-2}$, the uniform convergence is established. It is clear, therefore, that K converges uniformly on every compact subset of $|z|<(\sqrt{1+M}+\sqrt{M})^{-2}$. It follows from part (B) of Theorem 3.2 that $f(z)$ is analytic in this disk, and our proof is complete.

By further restricting the sequence $\{d_n\}$ in Theorem 3.3 we can say the following:

THEOREM 3.4. *Let* $\mathsf{K}=1-\delta_0 z+\mathsf{K}_{n=1}^{\infty}(d_nz/(1-\delta_nz))$ *be a* δ*-fraction such that*

$$\lim_{n\to\infty} d_n=0.$$

Then K *converges to a function* $f(z)$ *which is both meromorphic in the open unit disk* $D(0,1)=\{z:|z|<1\}$ *and analytic at* $z=0$. *The convergence is uniform on every compact subset of* $D(0,1)$ *which contains no poles of* $f(z)$.

Proof. By an equivalence transformation K can be put into the form

$$\mathsf{K}=1-\delta_0 z+\frac{d_1z/(1-\delta_1z)}{1}+\frac{\mathsf{K}_{n=2}^{\infty}(E_n(z)/1)}{1},$$

where

$$E_n(z)=\frac{d_nz}{(1-\delta_{n-1}z)(1-\delta_nz)},\qquad n\geq 2.$$

Clearly, each $E_n(z)$ is analytic if $|z|<1$. It is also easy to see that for each M satisfying $0<M<1$ there exists an n_M such that $|E_n(z)|\leq\frac{1}{4}$ for all $n\geq n_M$ and $|z|\leq M$. Thus, by the convergence neighborhood theorem referenced in the proof of Theorem 3.3, a tail

$$\mathsf{K}_{n=m}^{\infty}\left(\frac{E_n(z)}{1}\right)$$

of K converges uniformly to a function $F(z)$ on $|z|\leq M$ if $m\geq n_M$. By Theorem 3.2, $F(z)$ is analytic in the disk $|z|<M$ and $F(0)=0$. The remainder of the proof will not be given here since, after making the above observations, it is very similar to the proof suggested by Jones and Thron for [9, Thm. 7.23, p. 275].

The remaining theorems in this section deal with the convergence of δ-fractions having certain convergence criteria imposed on the sequences $\{d_n\}$ and $\{\delta_n\}$. As part of this investigation we introduce the following definition. We shall say that a δ-fraction

$$b_0-\delta_0z+\mathsf{K}_{n=1}^{\infty}\left(\frac{d_nz}{1-\delta_nz}\right)$$

is (p,q) *limit periodic* if there exist positive integers p and q such that

$$\lim_{\nu\to\infty} d_{p\nu+k}=D_k,\qquad k=0,1,\cdots,p-1$$

and

$$\lim_{\nu\to\infty}\delta_{q\nu+m}=\Delta_m, \qquad m=0,1,\cdots,q-1,$$

where each Δ_m is 0 or 1 and the D_k are numbers in the *extended* complex plane. Thus, in this setting, limit periodic regular C-fractions are in the class of $(1,1)$ limit periodic δ-fractions. It is not our intention in this paper to make a complete study of the convergence behavior of (p,q) limit periodic δ-fractions. However, as part of an introductory investigation illustrating some techniques that might be employed in a more complete study, we give in Theorems 3.5, 3.6 and 3.7 useful convergence criteria for the $(1,1)$, $(1,2)$ and $(2,1)$ cases, respectively. Before we state these theorems, it will be convenient for us throughout the remainder of this paper to introduce here the symbol $R[\alpha]$ for the *ray* from α to ∞ in the direction of α defined by

$$R[\alpha]=\{\alpha t: t\geq 1\}, \qquad \alpha\neq 0, \quad \alpha\in\mathbb{C}. \tag{3.9}$$

We note here also that in the next section we will give many examples of (p,q) limit periodic δ-fraction expansions for analytic functions.

THEOREM 3.5. *Let* $\mathsf{K}=1-\delta_0 z+\mathsf{K}_{n=1}^{\infty}(d_n z/(1-\delta_n z))$ *be a* δ*-fraction satisfying*

$$\lim_{n\to\infty} d_n=d, \qquad \lim_{n\to\infty}\delta_n=\delta,$$

where d *is a complex constant and* δ *is either* 0 *or* 1.

(A) *If* $d=\delta=0$, *then* K *converges to a function* $f(z)$ *which is both meromorphic in the complex plane* $\mathbb{C}$ *and analytic at* $z=0$. *The convergence is uniform on every compact subset of* $\mathbb{C}$ *which contains no poles of* $f(z)$.

(B) *If* $d=0$ *and* $\delta=1$, *then the conclusions are the same as in* (A) *with* $\mathbb{C}$ *replaced by the punctured plane* $\mathbb{C}-\{1\}$.

(C) *If* $d\neq 0$ *and* $\delta=0$, *then the conclusions are the same as in* (A) *with* $\mathbb{C}$ *replaced by the cut plane* $\mathbb{C}-R[-1/(4d)]$.

(D) *If* $d\neq 0$ *and* $\delta=1$, *then the conclusions are the same as in* (A) *with* $\mathbb{C}$ *replaced by any domain* D *such that* $0\in D$ *and* $D\subset\mathbb{C}-E_d$, *where*

$$E_d=\{z:[2-z-1/z]/(4d)\in[0,1]\}.$$

If d *is real, then* E_d *is a subset of the set made up of the real line and the unit circle. If* $d=1$, *in particular, then* E_d *is the unit circle.*

Proof. Let

$$E=\{z:1-\delta z=0\}\cup\left\{z:\frac{4dz}{(1-\delta z)^2}\in[-\infty,-1]\right\}$$

and let D be any domain (open connected set) in $\mathbb{C}$ satisfying

$$0\in D \quad\text{and}\quad D\cap E=\varnothing.$$

It is sufficient to prove the above theorem for the interior T^0 of T, where T is an arbitrary connected compact set such that $T\subset D$ and $0\in T^0$. Since D contains no points of E, it follows that the roots $x_1(z)$ and $x_2(z)$ of the quadratic equation

$$x^2-(1-\delta z)x-dz=0$$

have unequal moduli if $z\in D$. Thus, since $T\subset D$ and T is compact, there exist constants θ, C_1, and C_2 $(0<\theta<1$, $C_1>0$, and $C_2>0)$ such that

$$\frac{|x_2(z)|}{|x_1(z)|}\leq\theta \quad\text{and}\quad C_1\leq|x_1(z)|\leq C_2$$

for all $z \in T$. Hence, we have met the hypotheses of Perron [21, Satz 2.42, p. 93], so there exists an integer ν_0 such that the continued fraction

$$\mathsf{K}_\nu = 1 - \delta_\nu z + \mathop{\mathsf{K}}_{m=\nu+1}^{\infty} \left(\frac{d_m z}{1 - \delta_m z} \right)$$

converges uniformly on T if $\nu \geq \nu_0$. If $P_n(z)$ and $Q_n(z)$ denote the nth numerator and denominator, respectively, of K_ν, then from the determinant formula we have

$$\frac{P_{n+1}(z)}{Q_{n+1}(z)} - \frac{P_n(z)}{Q_n(z)} = \frac{(-1)^n z^{n+1} \Pi_{k=1}^{n+1} d_{\nu+k}}{Q_n(z) Q_{n+1}(z)}, \qquad n \geq 0.$$

Therefore,

$$\lambda\left(L\left(\frac{P_{n+1}}{Q_{n+1}} \right) - L\left(\frac{P_n}{Q_n} \right) \right) = n+1$$

and hence by Theorem 2.2, or by Jones and Thron [10, Thm. 1, p. 4], there exists a power series

$$L_\nu = 1 + C_1 z + C_2 z^2 + \cdots$$

to which K_ν corresponds at $z=0$, the order of correspondence being $n+1$. Each approximant of K_ν is a rational function analytic at the origin since $Q_n(0) \equiv 1$ for all $n \geq 0$.

Also, since K_ν converges uniformly on T, it converges uniformly on every compact subset of T^0 to a function $F_\nu(z)$. By [10, Thm. 4′, p. 15], $F_\nu(z)$ is analytic in T^0 and L is its Taylor series expansion about $z=0$. Now let A_k and B_k denote the kth numerator and denominator, respectively, of the original continued fraction K. Then for $n \geq 0$

$$\frac{A_{\nu+n}}{B_{\nu+n}} = \frac{A_{\nu-1}(P_n/Q_n) + z d_\nu A_{\nu-2}}{B_{\nu-1}(P_n/Q_n) + z d_\nu B_{\nu-2}}.$$

For $z \in T^0$ let

$$f(z) = \lim_{n\to\infty} \frac{A_{\nu+n}}{B_{\nu+n}} = \frac{A_{\nu-1} F_\nu(z) + z d_\nu A_{\nu-2}}{B_{\nu-1} F_\nu(z) + z d_\nu B_{\nu-2}}.$$

Since the numerator and denominator functions of the last expression for $f(z)$ have no common zeros and since the denominator is not identically zero in T^0 (it is equal to 1 at $z=0$), it follows that $f(z)$ is meromorphic in T^0. Using the facts that $\{P_n(z)/Q_n(z)\}$ converges uniformly to $F_\nu(z)$ on compact subsets of T^0, $0 \in T^0$ and that $B_{\nu-1}(z)F_\nu(z) + z d_\nu B_{\nu-2}(z)$ does not vanish at $z=0$, it is not difficult to verify that $\{A_k(z)/B_k(z)\}$ converges uniformly to $f(z)$ on pole free compact subsets of T^0. The fact that $f(z)$ is analytic at the origin follows from Theorem 3.3. After choosing d, δ, D and E appropriately for each of the four cases in the statement of the theorem, our proof is complete. Part (A) also follows from Jones and Thron [9, Thm. 7.23, p. 275] dealing with general T-fractions.

THEOREM 3.6. *Let* $\mathsf{K} - 1 - \delta_0 z + \mathsf{K}_{n=1}^{\infty}(d_n z/(1-\delta_n z))$ *be a* δ*-fraction satisfying*

$$\lim_{n\to\infty} d_n = d,$$

where d *is a complex constant, and either*

$$\lim_{n\to\infty} \delta_{2n} = 1, \qquad \lim \delta_{2n-1} = 0$$

or

$$\lim_{n\to\infty}\delta_{2n-1}=1, \qquad \lim_{n\to\infty}\delta_{2n}=0.$$

(A) *If* $d=0$, *then* K *converges to a function* $f(z)$ *which is both meromorphic in* $\mathbb{C}-\{1\}$ *and analytic at* $z=0$. *The convergence is uniform on every compact subset of* $\mathbb{C}-\{1\}$ *which contains no poles of* $f(z)$.

(B) *If* $d\neq 0$, *then the conclusions are the same as in* (A) *with* $\mathbb{C}-\{1\}$ *replaced by* $\mathbb{C}-E_d$, *where for real values of* d *the set* E_d *is given by*

(a) $\left[\frac{1}{1-4d},1\right]$ if $d<0$, (c) $[1,\infty)$ if $d=\frac{1}{4}$,

(b) $\left[1,\frac{1}{1-4d}\right]$ if $0<d<\frac{1}{4}$, (d) $\left(-\infty,-\frac{1}{4d-1}\right]\cup[1,\infty)$ if $d>\frac{1}{4}$,

and where, if $\mathrm{Im}(d)\neq 0$, E_d *is the arc of the circle with center* $1+id/2|\mathrm{Im}(d)|$ *and radius* $d/2|\mathrm{Im}(d)|$ *containing the three points* $1, 1/(1-4d)$ *and* $1/(1-d)$, *the first two as endpoints of the arc.*

Proof. By an equivalence transformation K can be put into the form

$$\mathsf{K}=1-\delta_0 z+\frac{d_1 z/(1-\delta_1 z)}{1}+\frac{d_2 z/(1-\delta_1 z)(1-\delta_2 z)}{1}+\frac{d_3 z/(1-\delta_2 z)(1-\delta_3 z)}{1}+\cdots$$

provided $z\neq 1$. Thus, in view of the conditions on the sequence $\{\delta_n\}$, for ν large enough a tail K_ν of K is of the form

$$\mathsf{K}_\nu=\frac{d_\nu w}{1}+\frac{d_{\nu+1}w}{1}+\frac{d_{\nu+2}w}{1}+\cdots,$$

where $w=z/(1-z)$. If $\lim_{n\to\infty} d_n=0$, then for each integer $M>1$ we have by [9, Thm. 4.55, p. 131] that there exists an integer ν such that K_ν converges to an analytic function of w for $|w|<M$, the convergence being uniform on each compact subset of $|w|<M$. But this implies that K_ν converges to a function $F_\nu(z)$ which is analytic for $z\in D_M$, where

$$D_M=\{z:|z-1/(1-M^{-2})|>M^{-1}/(1-M^{-2})\},$$

and that the convergence to $F_\nu(z)$ is uniform on compact subsets of D_M. It is easy to see that each compact subset of $\mathbb{C}-\{1\}$ is contained a region D_M for some $M>1$. Note, also, that $0\in D_M$ and $1\notin D_M$ for all $M>1$. Now that we have established the analyticity and uniform convergence of a tail of K on appropriate sets, the assertions in part (A) that K converges to a meromorphic function $f(z)$ in $\mathbb{C}-\{1\}$ (uniformly on pole free compact subsets) can be established by arguments similar to those used in the proofs of [9, Thm. 5.14, p. 182] and Theorem 3.5. The fact that $f(z)$ is analytic at $z=0$ follows from Theorem 3.3. This completes the proof of Part (A).

To prove part (B), where $\lim_{n\to\infty} d_n=d\neq 0$, we again look at the tail K_ν of K above for ν large enough. Let S be any compact connected subset of $\mathbb{C}-E_d$ such that S^0, the interior of S, contains the origin. Then, under the linear fractional transformation $w=z/(1-z)$, S maps onto a compact connected subset T of $\mathbb{C}-R[-1/(4d)]$ in the w-plane such that $w=0$ is in the interior T^0 of T. It is not difficult to verify that E_d is the image of the ray $R[-1/(4d)]$ under the inverse transformation $z=w/(1+w)$. By [9, Thm. 4.56, p. 132] the continued fraction

$$\mathsf{K}_\nu=\frac{d_\nu w}{w}+\frac{d_{\nu+1}w}{1}+\frac{d_{\nu+2}w}{1}+\cdots$$

converges in T^0 to an analytic function of w, and the convergence is uniform on compact subsets of T^0. It follows that K_ν, with $w=z/(1-z)$, converges in S^0 to a function $F_\nu(z)$ which is analytic in S^0 and that the convergence is uniform on compact subsets of S^0. Now that we have established that K has analytic tails on sets of the form S^0 if $d\neq 0$, the remainder of the proof of part (B) is similar to that suggested for part (A).

THEOREM 3.7. *Let*

$$\mathsf{K}=1-\delta_0 z+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{d_n z}{1-\delta_n z}\right)$$

be a δ-fraction satisfying

$$\lim_{n\to\infty} d_{2n}=\sigma_0,\quad \lim_{n\to\infty} d_{2n+1}=\sigma_1,\quad \lim_{n\to\infty}\delta_n=0,$$

where σ_0 and σ_1 are complex constants and $\sigma_0\neq\sigma_1$. Then K converges to a function $f(z)$ which is both meromorphic in $\mathbb{C}-E[\sigma_0,\sigma_1]$ and analytic at $z=0$, where

$$E[\sigma_0,\sigma_1]=E_1\cup E_2,$$

$$E_1=\left\{z=\frac{w/2}{\sqrt{\sigma_0\sigma_1}-(\sigma_0+\sigma_1)w/2}: w\in[-\infty,-1]\cup[1,\infty]\right\},$$

$$E_2=\{1\}\quad \text{if } \delta_n=1 \text{ for some } n\geq 1 \text{ and } \varnothing \text{ otherwise.}$$

The convergence is uniform on every compact subset of $\mathbb{C}-E[\sigma_0,\sigma_1]$ which contains no poles of $f(z)$. The "cut" $E[\sigma_0,\sigma_1]$ is a portion (bounded away from $z=0$) of a circle or a line passing through the origin plus possibly $z=1$ as an isolated point.

Proof. It is sufficient to prove the theorem for z in the interior T^0 of T, where $0\in T^0$ and T is an otherwise arbitrary compact connected subset of $\mathbb{C}-E[\sigma_0,\sigma_1]$. If $1-\delta_{2n}z\neq 0$ for all $n\geq 1$, then by [9, (2.4.24), p. 42] the even part of K is given by

$$b_0^{(0)}(z)+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{a_n^{(0)}(z)}{b_n^{(0)}(z)}\right), \tag{3.10}$$

where

$$a_1^{(0)}(z)=d_1 z(1-\delta_2 z),\quad a_2^{(0)}(z)=-d_2 d_3 z^2(1-\delta_4 z),$$

$$a_n^{(0)}(z)=-d_{2n-2}d_{2n-1}z^2(1-\delta_{2n-4}z)(1-\delta_{2n}z),\qquad n\geq 3,$$

$$b_0^{(0)}(z)=1-\delta_0 z, b_1^{(0)}(z)=(1-\delta_1 z)(1-\delta_2 z)+d_2 z,$$

$$b_n^{(0)}(z)=\left[(1-\delta_{2n-2}z)(1-\delta_{2n-1}z)+d_{2n-1}z\right](1-\delta_{2n}z)+d_{2n}z(1-\delta_{2n-2}z),\qquad n\geq 2.$$

If $1-\delta_{2n+1}z\neq 0$ for all $n\geq 1$, then by [9, (2.4.29), p. 43] the odd part of K is given by

$$b_0^{(1)}(z)+\mathop{\mathsf{K}}_{n=1}^{\infty}\left(\frac{a_n^{(1)}(z)}{b_n^{(1)}(z)}\right), \tag{3.11}$$

where

$$a_1^{(1)}(z)=-d_1 d_2 z^2(1-\delta_3 z)/(1-\delta_1 z),$$

$$a_n^{(1)}(z)=-d_{2n-1}d_{2n}z^2(1-\delta_{2n-3}z)(1-\delta_{2n+1}z),\qquad n\geq 2,$$

$$b_0^{(1)}(z)=\left[(1-\delta_0 z)(1-\delta_1 z)+d_1 z\right]/(1-\delta_1 z),$$

$$b_n^{(1)}(z)=\left[(1-\delta_{2n-1}z)(1-\delta_{2n}z)+d_{2n}z\right](1-\delta_{2n+1}z)+d_{2n+1}z(1-\delta_{2n-1}z),\qquad n\geq 1.$$

Since $\lim_{n\to\infty}\delta_n=0$, there exists an integer n_0 such that for all $n\geq n_0$

$$a_n^{(0)}(z)=-d_{2n-2}d_{2n-1}z^2, \qquad b_n^{(0)}(z)=1+(d_{2n-1}+d_{2n})z$$

and

$$a_n^{(1)}(z)=-d_{2n-1}d_{2n}z^2, \qquad b_n^{(1)}(z)=1+(d_{2n}+d_{2n+1})z.$$

Thus, for $z\in T$,

$$\lim_{n\to\infty} a_n^{(0)}(z)=\lim_{n\to\infty} a_n^{(1)}(z)=-\sigma_0\sigma_1 z^2$$

and

$$\lim_{n\to\infty} b_n^{(0)}(z)=\lim_{n\to\infty} b_n^{(1)}(z)=1+(\sigma_0+\sigma_1)z,$$

where in each case the convergence is uniform on T.

If $z\in\mathbb{C}-E[\sigma_0,\sigma_1]$, then $1+(\sigma_0+\sigma_1)z\neq 0$ and the roots $x_i(z)$ $(i=1,2)$ of the quadratic equation

$$x^2-[1+(\sigma_0+\sigma_1)z]x+\sigma_0\sigma_1 z^2=0$$

have unequal moduli, where

$$x_i(z)=\frac{1+(\sigma_0+\sigma_1)z}{2}\left[1-(-1)^i\sqrt{1-\frac{4\sigma_0\sigma_1 z^2}{(1+(\sigma_0+\sigma_1)z)^2}}\right], \qquad i=1,2.$$

The symbol $\sqrt{\ }$ denotes the square root with positive real part. Hence, since $T\subset\mathbb{C}-E[\sigma_0,\sigma_1]$ and T is compact, it follows, as in the proof of Theorem 3.5, that there exist positive constants θ, C_1, C_2 such that

$$C_1\leq|x_1(z)|\leq C_2, \qquad \frac{|x_2(z)|}{|x_1(z)|}\leq\theta<1.$$

Therefore, by [21, Satz 2.42, p. 93] there exists $\nu\geq n_0$ such that both

$$b_\nu^{(0)}+\mathop{\mathrm{K}}_{n=\nu+1}^{\infty}\left(\frac{a_n^{(0)}(z)}{b_n^{(0)}(z)}\right) \tag{3.12}$$

and

$$b_\nu^{(1)}+\mathop{\mathrm{K}}_{n=\nu+1}^{\infty}\left(\frac{a^{(1)}(z)}{b_n^{(1)}(z)}\right) \tag{3.13}$$

converge uniformly on T (and therefore on each compact subset of T^0) to functions $F_\nu^{(0)}(z)$ and $F_\nu^{(1)}(z)$, respectively. Let $R_n^{(0)}(z)$ and $R_n^{(1)}(z)$ denote the nth approximants of (3.12) and (3.13), respectively. Then

$$\lambda\big(R_{n+1}^{(i)}(z)-R_n^{(i)}(z)\big)=2n+2, \qquad i=0,1,$$

so by [10, Thm. 1, p. 4] each of the continued fractions (3.12) and (3.13) corresponds at $z=0$ to some power series of the form

$$L^{(i)}=1+c_1^{(i)}z+c_2^{(i)}z^2+\cdots, \qquad i=0,1,$$

respectively. Thus by [10, Thm. 4′, p. 15], the functions $F_\nu^{(i)}(z)$, $i=0,1$, are analytic in T^0 and $L^{(i)}$ is the Taylor series expansion of $F_\nu^{(i)}$ about $z=0$. Now that we have established the analyticity of the tails $F_\nu^{(i)}(z)$, arguments similar to those given in the

proof of Theorem 3.5 can be used to prove that the continued fractions (3.10) and (3.11) converge to meromorphic functions $f_v^{(i)}(z)$ $(i=1,2)$, respectively, in T^0, the convergence being uniform on compact subsets of T^0 containing no poles of the functions. Heavy use is made of the facts that $1-\delta_n z\neq 0$, $d_n\neq 0$, and the denominators of these continued fractions do not vanish at $z=0$. But the even and odd parts (3.10) and (3.11) of the original continued fraction K correspond at $z=0$ to the same power series L_0 that K does. Thus, since (by Theorem 3.3) K converges to an analytic function $f(z)$ in a neighborhood N of the origin and since T_0 contains such a neighborhood, it follows from Corollary 4.1 and [10, Thm. 4, p. 15] that $f_v^{(0)}(z)=f_v^{(1)}(z)=f(z)$ for $z\in N$, and f can be continued analytically in T^0 except for poles. This completes the proof of our theorem.

4. Expansions. We shall make repeated use in this section of the following known result on extension techniques for continued fractions. It is essentially Perron [21, Satz 1.7, p. 16], where a justification is also given.

THEOREM B. *If the approximants of the continued fraction*

$$b_0+\frac{a_1}{b_1}+\frac{a_2}{b_2}+\cdots \qquad (a_n\neq 0) \tag{4.1}$$

are A_n/B_n, then one can insert the approximant

$$\frac{A_k-\rho A_{k-1}}{B_k-\rho B_{k-1}} \qquad (\rho\neq 0)$$

between A_{k-1}/B_{k-1} and A_k/B_k and leave all others the same by replacing the section

$$\frac{a_k}{b_k}+\frac{a_{k+1}}{b_{k+1}}$$

of (4.1) *with the section*

$$\frac{a_k}{b_k-\rho}+\frac{\rho}{1}-\frac{a_{k+1}/\rho}{(b_{k+1}+a_{k+1}/\rho)}.$$

This modification can be applied infinitely often to obtain the following continued fraction

$$(b_0-\rho_0)+\frac{\rho_0}{1}-\frac{a_1/\rho_0}{(b_1+a_1/\rho_0-\rho_1)}+\frac{\rho_1}{1}-\frac{a_2/\rho_1}{(b_2+a_2/\rho_1-\rho_2)}+\frac{\rho_2}{1}-+\cdots \tag{4.2}$$

whose approximants, respectively, are

$$\frac{A_0-\rho_0}{B_0},\ \frac{A_0}{B_0},\ \frac{A_1-\rho_1A_0}{B_1-\rho_1B_0},\ \frac{A_1}{B_1},\ \frac{A_2-\rho_2A_1}{B_2-\rho_2B_1},\ \frac{A_2}{B_2},\ \ldots.$$

The continued fraction (4.2) *or any other continued fraction derived from* (4.1) *by the section changing or approximant insertion process described above is called an extension of* (4.1).

We shall also make heavy use of equivalence transformations of continued fractions in this section. For a thorough discussion of equivalence transformations, the reader is referred to Jones and Thron [9, §2.3] and to Perron [21, §2].

We are now ready to give, through a series of theorems, δ-fraction expansions for a variety of classical analytic functions as well as considerable information about the regions of validity of these expansions.

THEOREM 4.1. *The regular δ-fraction expansions of* $\tan z$ *and* $\tanh z$ *are given by*

$$\tan z = \frac{z}{1-z} + \frac{z}{1} + \frac{z/3}{1} - \frac{z/3}{1} - \frac{z/5}{1} + \frac{z/5}{1} + \frac{z/7}{1} - \frac{z/7}{1} - \frac{z/9}{1} + \cdots$$

and

$$\tanh z = \frac{z}{1-z} + \frac{z}{1} - \frac{z/3}{1} + \frac{z/3}{1} - \frac{z/5}{1} + \frac{z/5}{1} - \frac{z/7}{1} + \frac{z/7}{1} - \frac{z/9}{1} + \cdots.$$

These expansions are valid everywhere in the complex plane.

Proof. According to [12, p. 122], the following representation for $\tan z$ is valid everywhere in the complex plane except at points which are poles of the function

$$\tan z = \frac{z}{1} - \frac{z^2}{3} - \frac{z^2}{5} - \frac{z^2}{7} - \cdots - \frac{z^2}{2n+1} - \cdots. \tag{4.3}$$

By letting the ρ_n in Theorem B take on the values of z and $-z$ it is easily seen that the continued fraction (4.3) can be extended to the continued fraction

$$\frac{z}{1-z} + \frac{z}{1} + \frac{z}{3} - \frac{z}{1} - \frac{z}{5} + \frac{z}{1} + \frac{z}{7} - \frac{z}{1} - \frac{z}{9} + \frac{z}{1} + \cdots, \tag{4.4}$$

which by an equivalence transformation can be put into the form

$$\frac{z}{1-z} + \frac{z}{1} + \frac{z/3}{1} - \frac{z/3}{1} - \frac{z/5}{1} + \frac{z/5}{1} + \frac{z/7}{1} - \frac{z/7}{1} - \frac{z/9}{1} + \cdots. \tag{4.5}$$

The continued fraction (4.5) is a (1, 1) limit periodic δ-fraction satisfying

$$\lim_{n\to\infty} d_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \delta_n = 0.$$

Also, its $2n$th approximant is the nth approximant of (4.3). Hence, it follows from Theorem 3.5 that (4.5) converges to $\tan z$ everywhere in $\mathbb{C}$ except for poles.

From [12, p. 123] we obtain

$$\tanh z = \frac{z}{1} + \frac{z^2}{5} + \cdots + \frac{z^2}{2n+1} + \cdots, \tag{4.6}$$

valid everywhere in $\mathbb{C}$ except at the poles of $\tanh z$. By extension, (4.6) becomes

$$\frac{z}{1-z} + \frac{z}{1} - \frac{z}{3} + \frac{z}{1} - \frac{z}{5} + \frac{z}{1} - \frac{z}{7} + \frac{z}{1} - \cdots, \tag{4.7}$$

which is equivalent to

$$\frac{z}{1-z} + \frac{z}{1} - \frac{z/3}{1} + \frac{z/3}{1} - \frac{z/5}{1} + \frac{z/5}{1} - \frac{z/7}{1} + \frac{z/7}{1} - \cdots. \tag{4.8}$$

The even approximants of (4.8) are the approximants of (4.6), so again by Theorem 3.5, (4.8) converges to $\tanh z$ everywhere this function is defined.

THEOREM 4.2. (A) *If* $F(z)$ *is Dawson's integral function*

$$F(z) = e^{-z^2} \int_0^z e^{t^2}\, dt,$$

then the regular δ-fraction expansion of F, *valid everywhere in* $\mathbb{C}$, *is given by*

$$F(z) = \frac{z}{1-z} + \frac{z}{1} + \frac{d_3 z}{1} + \frac{d_4 z}{1} + \frac{d_5 z}{1} + \cdots, \tag{4.9}$$

where for $n \geq 1$

$$d_{4n-1} = -d_{4n} = \frac{(-1)^n n \binom{2n}{n}}{(4n-1)4^{n-1}}, \qquad d_{4n+2} = -d_{4n+1} = \frac{(-1)^n 4^n}{(4n+1)\binom{2n}{n}}, \tag{4.10}$$

and

$$|d_{4n-1}|\sim\frac{4\sqrt{n/\pi}}{4n-1},\qquad |d_{4n+2}|\sim\frac{\sqrt{n\pi}}{4n+1}. \tag{4.11}$$

(B) *If $E(z)$ is the modified error function defined by*

$$E(z)=\frac{\sqrt{\pi}}{2}e^{z^2}\operatorname{erf}(z)=e^{z^2}\int_0^z e^{-u^2}\,du,$$

then the regular δ-fraction expansion of E, valid everywhere in $\mathbb{C}$, is given by

$$E(z)=\frac{z}{1-z}+\frac{z}{1}+\frac{d_3^*z}{1}+\frac{d_4^*z}{1}+\frac{d_5^*z}{1}+\cdots, \tag{4.12}$$

where for $n\geq 1$

$$d_{4n}^*=-d_{4n-1}^*=d_{4n-1}\quad\text{and}\quad d_{4n+1}^*=-d_{4n+2}^*=d_{4n+2}.$$

(C) *If $C(z)$ and $S(z)$ denote the Fresnel integrals defined by*

$$C(z)=\int_0^z\cos(t^2)\,dt\quad\text{and}\quad S(z)=\int_0^z\sin(t^2)\,dt,$$

then the regular δ-fraction expansion of $e^{-iz^2}(C(z)+iS(z))$, valid everywhere in $\mathbb{C}$, is given by

$$e^{-iz^2}(C(z)+iS(z))=\frac{z}{1-z}+\frac{z}{1}+\frac{\hat{d}_3z}{1}+\frac{\hat{d}_4z}{1}+\frac{\hat{d}_5z}{1}+\cdots, \tag{4.13}$$

where for $n\geq 1$

$$\hat{d}_{4n-1}=-\hat{d}_{4n}=id_{4n-1}\quad\text{and}\quad\hat{d}_{4n+1}=-\hat{d}_{4n+2}=d_{4n+1}.$$

Proof. According to McCabe [16], the function $F(z)$ in part (A) can be represented by the continued fraction

$$F(z)=\frac{z}{1}+\frac{2z^2}{3}-\frac{4z^2}{5}+\frac{6z^2}{7}-\frac{8z^2}{9}+\cdots, \tag{4.14}$$

and the expansion is valid everywhere in the complex plane. We mention here, also, that Dijkstra [1] has given a certain continued fraction expansion for a generalization of Dawson's integral function. By an equivalence transformation, the continued fraction (4.14) can be put into the form

$$F(z)=\frac{z}{1}+\frac{2z^2/3}{1}-\frac{4x^2/3\times 5}{1}+\frac{6z^2/5\times 7}{1}-\frac{8z^2/7\times 9}{1}+-\cdots. \tag{4.15}$$

By extending (4.15) we obtain the δ-fraction

(4.16)

$$\begin{aligned}&\frac{z}{1-z}+\frac{z}{1}-\frac{2z/3}{1}+\frac{2z/3}{1}+\frac{(4/2)(z/5)}{1}-\frac{(4/2)(z/5)}{1}+\frac{(2\times 6)(z/7)}{1}-\frac{(2\times 6)/4(z/7)}{1}\\&\quad-\frac{(4\times 8/2\times 6)(z/9)}{1}+\frac{(4\times 8/2\times 6)(z/9)}{1}-\frac{(2\times 6\times 10/4\times 8)(z/11)}{1}\\&\quad+\frac{(2\times 6\times 10/4\times 8)(z/11)}{1}+\frac{(4\times 8\times 12/2\times 6\times 10)(z/13)}{1}-\cdots.\end{aligned}$$

It can be shown that (4.16) is the same as the continued fraction (4.9), where the d_n are given by (4.10). It follows from the asymptotic formulas (4.11) (which were determined with the aid of Stirling's formula) that

$$\lim_{n\to\infty} d_n = 0.$$

Also, the $2n$th ($n \geq 0$) approximant of the continued fraction (4.9) is the nth approximant of (4.15). Thus it follows from Theorem 3.5 that (4.9) converges to $F(z)$ at all points $z \in \mathbb{C}$, and our proof of part (A) is complete.

The expansion for $E(z)$ in part (B) is derived in a manner similar to that for $F(z)$ in part (A). A simple change of variables computation will show that

$$E(z) = -iF(iz)$$

so, using the expansion (4.15) for $F(z)$, we obtain

$$(4.17)\qquad E(z) = \frac{z}{1} - \frac{2z^2/1\times 3}{1} + \frac{4z^2/3\times 5}{1} - \frac{6z^2/5\times 7}{1} + \frac{8z^2/7\times 9}{1} - + \cdots.$$

The δ-fraction expansion (4.12) for $E(z)$ given in part (B) of our Theorem is an extension of (4.17), and its even approximants are the approximants of (4.17).

To prove part (C) we make use of the following known (see [9, p. 209]) continued fraction expansion for the confluent hypergeometric function $\Phi(1; c; z)$, where $c \notin \{0, -1, -2, -3, \cdots\}$:

$$(4.18)\qquad \Phi(1;c;z) = \frac{1}{1} - \frac{z/c}{1} + \frac{z/c(c+1)}{1} - \frac{cz/(c+1)(c+2)}{1} + \frac{2z/(c+2)(c+3)}{1}$$

$$- \frac{(c+1)z/(c+3)(c+4)}{1} + \frac{3z/(c+4)(c+5)}{1} - \cdots.$$

By making a change of variables in [9, (6.1.41), p. 208] it is easily verified that

$$C(z) + iS(z) = ze^{iz^2}\Phi\left(1; \frac{3}{2}; -iz^2\right).$$

Using this formula and the continued fraction (4.18), which represents $\Phi(1; c; z)$ for all $z \in \mathbb{C}$, we obtain

(4.19)

$$e^{-iz^2}(C(z) + iS(z))$$

$$= \frac{z}{1} + \frac{2iz^2/3}{1} - \frac{(2/3\times 2/5)(iz^2)}{1} + \frac{(2/5\times 2/7)(3iz^2/2)}{1}$$

$$- \frac{(2/7\times 2/9)(2iz^2)}{1} + \frac{(2/9\times 2/11)(5iz^2/2)}{1} - \frac{(2/11\times 2/13)(3iz^2)}{1} + \cdots.$$

The continued fraction (4.13) is an extension of (4.19), and the $2n$th approximant of (4.13) is the nth approximant of (4.19). Since the coefficients $\hat{d}_n$ in (4.13) tend to zero as $n \to \infty$, it now follows from Theorem 3.5 that this continued fraction converges everywhere to the function $e^{-iz^2}(C(z) + iS(z))$. With this our proof is complete.

THEOREM 4.3. *If J_n denotes the Bessel function of nth order, then the regular δ-fraction expansion of J_m/J_{m-1}, $m\geq 1$, is given by*

(4.20)

$$\frac{J_m(z)}{J_{m-1}(z)}=\frac{z/2m}{1-z}+\frac{z}{1}+\frac{z/4m(m+1)}{1}-\frac{z/4m(m+1)}{1}-\frac{mz/(m+2)}{1}+\frac{mz/(m+2)}{1}$$
$$+\frac{z/4m(m+3)}{1}-\frac{z/4m(m+3)}{1}-\frac{mz/(m+4)}{1}+\frac{mz/(m+4)}{1}+\cdots$$
$$=\frac{d_1z}{1-z}+\frac{d_2z}{1}+\frac{d_3z}{1}+\frac{d_4z}{1}+\cdots,$$

where $d_1=1/2m$, $d_2=1$ and (for $n\geq 1$)

$$d_{4n-1}=-d_{4n}=\frac{1}{4m(m+2n-1)},\qquad d_{4n+2}=-d_{4n+1}=\frac{m}{m+2n}. \tag{4.21}$$

This expansion is valid everywhere in the complex plane.

Proof. From Khovanskii [12, p. 133] we obtain

$$\frac{J_m(z)}{J_{m-1}(z)}=\frac{z/2m}{1}-\frac{z^2/4m(m+1)}{1}-\cdots-\frac{z^2/4(m+n)(m+n-1)}{1}-\cdots, \tag{4.22}$$

and the expansion is valid for all $z\in\mathbb{C}$. The continued fraction (4.20) is an extension of the continued fraction (4.22). If $g_n(z)$ denotes the nth approximant of (4.20) and if $A_n(z)$ ($B_n(z)$) denotes the numerator (denominator) of the nth approximant of (4.22), then it is not difficult to verify the relations

$$g_{2n}(z)=\frac{A_n(z)}{B_n(z)}\quad\text{and}\quad g_{2n-1}(z)=\frac{A_n(z)-\rho_n zA_{n-1}(z)}{B_n(z)-\rho_n zB_{n-1}(z)},$$

where

$$\rho_{2n}=-\frac{1}{4m(m+2n-1)},\qquad \rho_{2n-1}=\frac{m}{m+2n-2},\qquad n\geq 1.$$

We have $g_{2n}(z)=A_n(z)/B_n(z)\to J_m(z)/J_{m-1}(z)$ everywhere as $n\to\infty$. Also, the coefficients d_n in (4.20) have the property that $\lim_{n\to\infty}d_n=0$. This along with Theorem 3.5 guarantees the validity of (4.20).

The expansion (4.23) for $\log(1+z)$ in our next theorem is just an equivalent version of a well-known expansion (see, for example, [12, eq. 4.2, p. 110]). It is included here for the sake of comparison and completeness and to help derive the new δ-fraction expansion for $\log(1+z^2)$. The δ-fraction (4.24) has an interesting feature in that it is (4, 1) limit periodic with $\lim_{n\to\infty}d_{4n-1}=\lim_{n\to\infty}d_{4n-2}=0$ and $\lim_{n\to\infty}d_{4n+1}=\lim_{n\to\infty}d_{4n}=\infty$. In the proof of Theorem 4.4 we employ for the first time a technique, based on Poincaré's theorem, for establishing the convergence behavior of (4.24).

THEOREM 4.4. (A) *The regular δ-fraction expansion for $\log(1+z)$ is given by*

$$\log(1+z)=\frac{z}{1}+\frac{z/1\times 2}{1}+\frac{z/2\times 3}{1}+\frac{2z/2\times 3}{1}+\frac{2z/2\times 5}{1}+\frac{3z/2\times 5}{1}$$
$$+\frac{3z/2\times 7}{1}+\cdots+\frac{nz/2(2n-1)}{1}+\frac{nz/2(2n+1)}{1}+\cdots \tag{4.23}$$

The expansion is valid for all $z\in\mathbb{C}-R[-1]$.

(B) *The regular δ-fraction expansion for* $\log(1+z^2)$ *is given by*

$$\begin{aligned}(4.24)\quad \log(1+z^2) &= -z+\frac{z}{1}-\frac{z/1}{1}+\frac{z/1}{1}-\frac{z/2}{1}+\frac{z/2}{1}-\frac{z/3}{1}+\frac{z/3}{1}\\ &\quad -\frac{2z/2}{1}+\frac{2z/2}{1}-\frac{z/5}{1}+\frac{z/5}{1}-\frac{3z/2}{1}+\frac{3z/2}{1}-\frac{z/7}{1}+\frac{z/7}{1}-\cdots\\ &= -z+\frac{z}{1+}\frac{d_2z}{1+}\frac{d_3z}{1+}\frac{d_4z}{1+}\cdots,\end{aligned}$$

where

$$(4.25)\qquad d_{4n-1}=-d_{4n-2}=\frac{1}{2n-1},\qquad d_{4n+1}=-d_{4n}=\frac{n}{2},\qquad n\geq 1.$$

The expansion (4.24) *is valid for all* $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$.

Proof. From expansion (4.23) we obtain

(4.26)

$$\begin{aligned}\log(1+z^2) &= \frac{z^2}{1}+\frac{z^2/1\times 2}{1}+\frac{z^2/2\times 3}{1}+\frac{2z^2/2\times 3}{1}+\frac{2z^2/2\times 5}{1}\\ &\quad +\frac{3z^2/2\times 5}{1}+\frac{3z^2/2\times 7}{1}+\cdots+\frac{nz^2/2(2n-1)}{1}+\frac{nz^2/2(2n+1)}{1}+\cdots,\end{aligned}$$

valid if $z^2\in\mathbb{C}-R[-1]$ or, equivalently, if $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$. By an equivalence transformation, (4.26) can be put into the form

$$\begin{aligned}(4.27)\quad \log(1+z^2) &= \frac{z^2}{1}+\frac{z^2}{(2/1)}+\frac{z^2}{3}+\frac{z^2}{(2/2)}+\frac{z^2}{5}+\frac{z^2}{(2/3)}\\ &\quad +\frac{z^2}{7}+\frac{z^2}{(2/4)}+\frac{z^2}{9}+\frac{z^2}{(2/5)}+\frac{z^2}{11}+\cdots\\ &= \frac{z^2}{b_1}+\frac{z^2}{b_2}+\frac{z^2}{b_3}+\cdots,\end{aligned}$$

where $b_{2n-1}=2n-1$ and $b_{2n}=2/n$. We extend (4.27), using $\rho_n\equiv z$ in Theorem B, to obtain the continued fraction

$$(4.28)\qquad -z+\frac{z}{1}-\frac{z}{b_1}+\frac{z}{1}-\frac{z}{b_2}+\frac{z}{1}-\frac{z}{b_3}+\frac{z}{1}-\cdots,$$

where the b_n are given above. The δ-fraction (4.24) is now easily obtained from (4.28) by an equivalence relation. It remains to investigate the convergence of (4.24). Let $A_n(z)(B_n(z))$ denote the nth numerator (denominator) of the continued fraction (4.26), let $f_n(z)=A_n(z)/B_n(z)$, and let $g_n(z)$ denote the nth approximant of (4.24). Then

$$g_{2n+1}(z)=f_n(z)\quad\text{and}\quad g_{2n}(z)=\frac{A_n(z)-(z/b_n)A_{n-1}(z)}{B_n(z)-(z/b_n)B_{n-1}(z)},\qquad n\geq 0.$$

Hence, for $n\geq 0$,

$$(4.29)\qquad |g_{4n+2}(z)-f_{2n+1}(z)|=\frac{|z/b_{2n+1}||f_{2n+1}(z)-f_{2n}(z)|}{|(B_{2n+1}(z)/B_{2n}(z))-z/b_{2n+1}|}$$

and

$$(4.30)\qquad |g_{4n}(z)-f_{2n}(z)|=\frac{|z||f_{2n}(z)-f_{2n-1}(z)|}{|(b_{2n}B_{2n}(z)/B_{2n-1}(z))-z|}.$$

Let $a_n(z)$, $n\geq 1$, denote the nth partial numerator in (4.26). Since (4.26) is limit periodic with $\lim_{n\to\infty} a_n(z)=z^2/4$ and since $B_n(z)=B_{n-1}(z)+a_n(z)B_{n-2}(z)$, it follows from Theorem A that $\{B_n(z)/B_{n-1}(z)\}$ must converge to one of the two roots $x_1(z)$ and $x_2(z)$ of the equation

$$x^2-x-\frac{z^2}{4}=0,$$

provided

$$|x_1(z)|\neq|x_2(z)|.$$

It is easily verified that $|x_1(z)|\neq|x_2(z)|$ if $z^2\in\mathbb{C}-R[-1]$. In particular, if z is a nonzero real number, then $\lim_{n\to\infty} B_n(z)/B_{n-1}(z)=(1+\sqrt{1+z^2})/2$, since the other choice is negative and therefore not possible. Clearly, (4.24) converges to 0 if $z=0$, so let us assume now that $z\neq 0$. Thus, if $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$ and noting that $b_{2n+1}\to\infty$ and $b_{2n}\to 0$ as $n\to\infty$, it follows from formulas (4.29) and (4.30) that the expansion (4.24) converges to $\log(1+z^2)$. This completes our proof of Theorem 4.4.

The expansion (4.31) in our next Theorem is easily derived from a well-known continued fraction expansion originally due to Lagrange (see [12, p. 102] or [21, p. 152]). Thus our proof of Theorem 4.5 shall concentrate on justifying the new δ-fraction expansions (4.33) and (4.34) over the region indicated. Murphy [20] has given various continued fraction expansions for $(1+z^2)^{-1/2}$. We deal with the functions $(1+z^2)^{\nu}$ for all ν satisfying $0<\nu<1$.

THEOREM 4.5. (A) *Suppose ν is any real number satisfying $0<\nu<1$. Then the regular δ-fraction expansion of $(1+z)^{\nu}$, valid for all $z\in\mathbb{C}-R[-1]$, is given by*

(4.31)

$$(1+z)^{\nu}=1+\frac{\nu z}{1}+\frac{(1-\nu)z/2}{1}+\frac{(1+\nu)/6}{1}+\frac{(2-\nu)z/6}{1}+\frac{(2+\nu)z/10}{1}$$
$$+\frac{(3-\nu)z/10}{1}+\cdots+\frac{(n+\nu)z/(4n+2)}{1}+\frac{(n+1-\nu)z/(4n+2)}{1}+\cdots.$$

In particular,

$$\sqrt{1+z}=1+\frac{z/2}{1}+\frac{z/4}{1}+\frac{z/4}{1}+\frac{z/4}{1}+\cdots. \tag{4.32}$$

(B) *The regular δ-fraction expansion of $(1+z^2)^{\nu}$ $(0<\nu<1)$, valid for all $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$, is given by*

$$(1+z^2)^{\nu}=(1-z)+\frac{z}{1}-\frac{c_1z}{1}+\frac{c_1z}{1}-\frac{c_2z}{1}+\frac{c_2z}{1}-\frac{c_3z}{1}+\frac{c_3z}{1}-+\cdots, \tag{4.33}$$

where, if $B(x,y)$ denotes the beta function and $n\geq 1$,

$$c_{2n-1}=\left(\frac{n}{2n-1}\right)\frac{B(n+\nu,1-\nu)}{B(n-\nu,\nu)}\to\begin{cases}0 & \text{if } 0<\nu<\frac{1}{2}\\ \infty & \text{if } \frac{1}{2}<\nu<1\end{cases}\quad \text{as } n\to\infty,$$

$$c_{2n}=\left(\frac{1}{2}\right)\frac{B(n+1-\nu,\nu)}{B(n+\nu,1-\nu)}\to\begin{cases}\infty & \text{if } 0<\nu<\frac{1}{2}\\ 0 & \text{if } \frac{1}{2}<\nu<1\end{cases}\quad \text{as } n\to\infty.$$

In particular,

$$\sqrt{1+z^2}=(1-z)+\cfrac{z}{1-\cfrac{z/2}{1+\cfrac{z/2}{1-\cfrac{z/2}{1+\cfrac{z/2}{1-+\cdots}}}}}. \tag{4.34}$$

Proof. After substituting z^2 for z in (4.31) we obtain

(4.35)

$$(1+z^2)^{\nu}=1+\cfrac{\nu z^2}{1}+\cfrac{(1-\nu)z^2/2}{1}+\cfrac{(1+\nu)z^2/6}{1}+\cfrac{(2-\nu)z^2/6}{1}+\cfrac{(2+\nu)z^2/10}{1}$$
$$+\cfrac{(3-\nu)z^2/10}{1}+\cdots+\cfrac{(n+\nu)z^2/(4n+2)}{1}+\cfrac{(n+1-\nu)z^2/(4n+2)}{1}+\cdots.$$

By an equivalence transformation, this continued fraction can be put into the form

$$1+\cfrac{z^2}{b_1+\cfrac{z^2}{b_2+\cfrac{z^2}{b_3+\cdots}}}, \tag{4.36}$$

where $b_1=1/\nu$ and

$$b_{2n+1}=\frac{(2n+1)(1-\nu)\cdots(n-\nu)}{\nu(1+\nu)\cdots(n+\nu)}, \qquad b_{2n}=\frac{2\nu(1+\nu)\cdots(n-1+\nu)}{(1-\nu)(2-\nu)\cdots(n-\nu)} \qquad (n\geq 1).$$

With the aid of the functional relations

$$\Gamma(x+1)=x\Gamma(x) \quad \text{and} \quad B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where $\Gamma(x)$ is the gamma function, the following formulas for the b_n can be obtained:

$$b_{2n-1}=\left(\frac{2n-1}{n}\right)\frac{B(n-\nu,\nu)}{B(n+\nu,1-\nu)}, \qquad b_{2n}=\frac{2B(n+\nu,1-\nu)}{B(n+1-\nu,\nu)} \qquad (n\geq 1). \tag{4.37}$$

Using Theorem B with $\rho_n\equiv z$ $(n\geq 0)$, we extend (4.36) to obtain

$$(1-z)+\cfrac{z}{1-\cfrac{z}{b_1+\cfrac{z}{1-\cfrac{z}{b_2+\cfrac{z}{1-\cfrac{z}{b_3+\cfrac{z}{1-\cdots}}}}}}}, \tag{4.38}$$

which, after an equivalence transformation and after setting $c_n=1/b_n$ $(n\geq 1)$, becomes (4.33). Let $A_n(z)(B_n(z))$ denote the nth numerator (denominator) of (4.35), $f_n(z)=A_n(z)/B_n(z)$, $h_n(z)=B_n(z)/B_{n-1}(z)$, and let $g_n(z)$ denote the nth approximant of (4.33). Then, for all $n\geq 0$,

$$g_{2n+1}(z)=f_n(z) \quad \text{and} \quad g_{2n}(z)=\frac{A_n(z)-zc_nA_{n-1}(z)}{B_n(z)-zc_nB_{n-1}(z)}.$$

Hence,

$$|g_{4n}(z)-f_{2n}(z)|=\frac{|zc_{2n}||f_{2n}(z)-f_{2n-1}(z)|}{|h_{2n+1}(z)-zc_{2n}|} \tag{4.39}$$

and

$$|g_{4n+2}(z)-f_{2n+1}(z)|=\frac{|zc_{2n+1}||f_{2n+1}(z)-f_{2n}(z)|}{|h_{2n+1}(z)-zc_{2n+1}|}. \tag{4.40}$$

The continued fraction (4.35) is limit periodic because its partial numerators converge to $z^2/4$, and the roots $x_i(z)$ $(i=1,2)$ of the associated quadratic equation

$$x^2-x-\frac{z^2}{4}=0$$

have unequal moduli if $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$. Therefore, it follows from Theorem A that $\{h_n(z)\}$ must converge to $x_1(z)$ or $x_2(z)$ if z is in this region. We have that $c_n\equiv\frac{1}{2}$ if $\nu=\frac{1}{2}$, and with the aid of Stirling's formula it can be verified that

$$\lim_{n\to\infty} c_{2n-1}=0(\infty) \text{ if } 0<\nu<\frac{1}{2} \qquad \left(\frac{1}{2}<\nu<1\right)$$

and

$$\lim_{n\to\infty} c_{2n}=\infty(0) \text{ if } 0<\nu<\frac{1}{2} \qquad \left(\frac{1}{2}<\nu<1\right).$$

We are now in a position to establish the convergence behavior of (4.33) asserted in part (B). By convention, the continued fraction (4.33) converges to 1 if $z=0$, so let us now assume that $z\neq 0$ and $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$. Then, using the convergence properties of $\{c_n\}$ and $\{h_n(z)\}$ established above, it follows that the right sides of (4.39) and (4.40) converge to 0 as $n\to\infty$. Hence, since $\lim_{n\to\infty} f_n(z)=(1+z^2)^\nu$, it follows that

$$\lim_{n\to\infty} g_{4n}(z)=\lim_{n\to\infty} g_{4n+2}(z)=\lim_{n\to\infty} g_{2n+1}(z)=(1+z^2)^\nu$$

and our proof is complete.

THEOREM 4.6. (A) *The regular δ-fraction expansion of* $(\arcsin z)/\sqrt{1-z^2}$, *valid for all* $z\in\mathbb{C}-\{R[-1]\cup R[1]\}$, *is given by*

(4.41)

$$\frac{\arcsin z}{\sqrt{1-z^2}}=\frac{z}{1-z}+\frac{z}{1}+\frac{(1\times 2)z/3}{1}-\frac{(1\times 2)z/3}{1}-\frac{z/5}{1}+\frac{z/5}{1}$$
$$+\frac{(3\times 4)z/7}{1}-\frac{(3\times 4)z/7}{1}-\frac{z/9}{1}+\frac{z/9}{1}+\frac{(5\times 6)z/11}{1}-\frac{(5\times 6)z/11}{1}-\cdots$$
$$=\frac{d_1}{1-z}+\frac{d_2z}{1}+\frac{d_3z}{1}+\cdots,$$

where $d_1=d_2=1$ *and*

$$d_{4n-1}=-d_{4n}=\frac{(2n-1)(2n)}{4n-1}, \qquad d_{4n+2}=-d_{4n+1}=\frac{1}{4n+1} \qquad (n\geq 1).$$

(B) *The regular δ-fraction expansion of* $(\operatorname{arcsinh} z)/\sqrt{1+z^2}$, *valid for all* $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$, *is given by*

(4.42)

$$\frac{\operatorname{arcsinh} z}{\sqrt{1+z^2}}=\frac{z}{1-z}+\frac{z}{1}-\frac{(1\times 2)z/3}{1}+\frac{(1\times 2)z/3}{1}-\frac{z/5}{1}+\frac{z/5}{1}$$
$$-\frac{(3\times 4)z/7}{1}+\frac{(3\times 4)z/7}{1}-\frac{z/9}{1}+\frac{z/9}{1}-\frac{(5\times 6)z/11}{1}+\frac{(5\times 6)z/11}{1}-\cdots$$
$$=\frac{d_1}{1-z}+\frac{d_2z}{1}+\frac{d_3z}{1}+\cdots,$$

where $d_1=d_2=1$ *and*

$$d_{4n}=-d_{4n-1}=\frac{(2n-1)(2n)}{4n-1}, \qquad d_{4n+2}=-d_{4n+1}=\frac{1}{4n+1} \qquad (n\geq 1).$$

Proof. From [9, formula 6.1.20, p. 203] we obtain

$$\frac{\arcsin z}{\sqrt{1-z^2}}=\frac{z}{1}-\frac{1\times 2z^2}{3}-\frac{1\times 2z^2}{5}-\frac{3\times 4z^2}{7}-\frac{3\times 4z^2}{9}-\frac{5\times 6z^2}{11}-\cdots, \tag{4.43}$$

valid for all $z\in\mathbb{C}-\{R[-1]\cup R[1]\}$. The continued fraction (4.43) is equivalent to

$$\frac{\arcsin z}{\sqrt{1-z^2}}=\frac{z}{1}-\frac{1\times 2z^2/1\times 3}{1}-\frac{1\times 2z^2/3\times 5}{1}-\frac{3\times 4z^2/5\times 7}{1}-\frac{3\times 4z^2/7\times 9}{1}$$
$$-\frac{5\times 6z^2/9\times 11}{1}-\frac{5\times 6z^2/11\times 13}{1}-\cdots. \tag{4.44}$$

The continued fraction (4.41) is obtained by extending (4.44). If $f_n(z)=A_n(z)/B_n(z)$ and $g_n(z)$ denote the nth approximants of (4.44) and (4.41), respectively, and if $h_n(z)=B_n(z)/B_{n-1}(z)$, then

$$g_{2n}(z)=f_n(z) \quad \text{and} \quad g_{2n-1}(z)=\frac{A_n(z)-zc_nA_{n-1}(z)}{B_n(z)-zc_nB_{n-1}(z)}, \tag{4.45}$$

where

$$c_{2n}=-\frac{(2n-1)(2n)}{4n-1}, \quad n\geq 1, \qquad c_{2n+1}=\frac{1}{4n+1}, \quad n\geq 0.$$

From (4.45) we derive

$$|g_{4n-1}(z)-f_{2n}(z)|=\frac{|z||f_{2n}(z)-f_{2n-1}(z)|}{|(h_{2n}(z)/c_{2n})-z|} \tag{4.46}$$

and

$$|g_{4n+1}(z)-f_{2n+1}(z)|=\frac{|zc_{2n+1}||f_{2n+1}(z)-f_{2n}(z)|}{|h_{2n+1}(z)-zc_{2n+1}|}. \tag{4.47}$$

Since the partial numerators of (4.44) converge to $-z^2/4$, it follows from Theorem A that $\{h_n(z)\}$ converges to one of the roots of

$$x^2-x+\frac{z^2}{4}=0$$

if $z\in\mathbb{C}-\{R[-1]\cup R[1]\}$. By an argument similar to that given in the proof of Theorem 4.5, it follows that the right sides of (4.46) and (4.47) tend to zero as $n\to\infty$ if $z\neq 0$. So if z is in the above region and $z\neq 0$, it follows that

$$\lim_{n\to\infty} g_{4n-1}(z)=\lim_{n\to\infty} g_{4n+1}(z)=\lim_{n\to\infty} g_{2n}(z)=\lim_{n\to\infty} f_n(z)=\frac{\arcsin z}{\sqrt{1-z^2}}.$$

Clearly, (4.43) is valid when $z=0$. This completes our argument for part (A).

Since the proof of part (B) parallels the proof for part (A) we shall not present it here. However, we do give below the key continued fractions involved in the derivation

of (B). From [12, p. 121] we obtain

$$\frac{\operatorname{arcsinh} z}{\sqrt{1+z^2}} = \frac{z}{1} + \frac{1\times 2z^2}{3} + \frac{1\times 2z^2}{5} + \cdots + \frac{(2n-1)(2n)z^2}{4n-1} + \frac{(2n-1)(2n)z^2}{4n+1} + \cdots, \tag{4.48}$$

which after an equivalence transformation becomes

$$\frac{\operatorname{arcsinh} z}{\sqrt{1+z^2}} = \frac{z}{1} + \frac{1\times 2z^2/1\times 3}{1} + \frac{1\times 2z^2/3\times 5}{1} + \frac{3\times 4z^2/5\times 7}{1} + \frac{3\times 4z^2/7\times 9}{1} + \cdots. \tag{4.49}$$

The continued fraction (4.42) is an extension of (4.49), and it has the property that its $2n$th approximant is the nth approximant of (4.49).

We feel that the continued fraction expansions given in our next theorem are especially interesting. This is partly because they are examples of convergent (4, 1) limit periodic δ-fractions having the property that the four subsequential limits $\pm 1/\pi$ and $\pm\pi/4$ of the sequence $\{d_n\}$ are finite and transcendental.

THEOREM 4.7. (A) *The regular δ-fraction expansion of* $\arctan z$ *is given by*

$$\arctan z = \frac{z}{1-z} + \frac{z}{1} + \frac{d_3 z}{1} + \frac{d_4 z}{1} + \frac{d_5 z}{1} + \cdots, \tag{4.50}$$

where (for $n \geq 1$)

$$\begin{aligned} d_{4n} &= -d_{4n-1} = \frac{n^2\left[\binom{2n}{n}\right]^2}{(4n-1)4^{2n-1}} \sim \frac{4n}{\pi(4n-1)} \to \frac{1}{\pi} \quad \text{as } n\to\infty, \\ d_{4n+2} &= -d_{4n+1} = \frac{4^{2n}}{(4n+1)\left[\binom{2n}{n}\right]^2} \sim \frac{\pi n}{4n+1} \to \frac{\pi}{4} \quad \text{as } n\to\infty. \end{aligned} \tag{4.51}$$

This expansion is valid for all $z \in \mathbb{C} - \{R[-i] \cup R[i]\}$ except possibly at $z = (\pi/4 - 1/\pi)^{-1}$.

(B) *The regular δ-fraction expansion of* $\operatorname{arctanh} z$ *is given by*

$$\operatorname{arctanh} z = \frac{z}{1-z} + \frac{z}{1} + \frac{c_3 z}{1} + \frac{c_4 z}{1} + \frac{c_5 z}{1} + \cdots, \tag{4.52}$$

where (for $n \geq 1$)

$$c_{4n-1} = -c_{4n} = d_{4n}, \qquad c_{4n+2} = -c_{4n+1} = d_{4n+2}$$

with d_{4n} and d_{4n+2} defined by (4.51). *This expansion is valid for all $z \in \mathbb{C} - \{R[-1] \cup R[1]\}$ except possibly at $z = (\pi/4 + 1/\pi)^{-1}$.*

Proof. The following known representation of $\arctan z$ is taken from [9, p. 202]:

$$\arctan z = \frac{z}{1} + \frac{1^2 z^2}{3} + \frac{2^2 z^2}{5} + \frac{3^2 z^2}{7} + \frac{4^2 z^2}{9} + \cdots, \tag{4.53}$$

valid if $z \in \mathbb{C} - \{R[-i] \cup R[i]\}$. The continued fraction (4.53) is equivalent to

$$\arctan z = \frac{z}{1} + \frac{1^2 z^2/1\times 3}{1} + \frac{2^2 z^2/3\times 5}{1} + \frac{3^2 z^2/5\times 7}{1} + \frac{4^2 z^2/7\times 9}{1} + \cdots. \tag{4.54}$$

The continued fraction (4.50) is derived as an extension of (4.54) as follows: Let $a_n(z)$, $n=1,2,\cdots$, denote the nth partial numerators of (4.54). Then (4.50) is identical to

$$\frac{a_1(z)}{1-\rho_1(z)+}\ \frac{\rho_1(z)}{1}\ \frac{a_2(z)/\rho_1(z)}{-1-\rho_2(z)+a_2(z)/\rho_1(z)+}\ \frac{\rho_2(z)}{1}\ \frac{a_3(z)/\rho_2(z)}{-1-\rho_3(z)+a_3(z)/\rho_2(z)+}\cdots,$$

where $\{\rho_n(z)\}$ is defined by

$$\rho_1(z)=z, \qquad \rho_{n+1}(z)=\frac{a_{n+1}(z)}{\rho_n(z)}, \qquad n\geq 1.$$

The formulas (4.51) for the coefficients $\{d_n\}$ of (4.50) and their asymptotic behavior can be derived from the above process with a modest amount of algebraic manipulation and Stirling's formula. The continued fraction (4.50) certainly converges to $\arctan z$ when $z=0$, so from now on in our convergence investigation we assume $z\neq 0$. Let $f_n(z)=A_n(z)/B_n(z)$ denote the nth approximant of (4.54), $g_n(z)$ the same for (4.50), and $h_n(z)=B_n(z)/B_{n-1}(z)$. Then $g_{2n}(z)=f_n(z)$, $n=0,1,2,\cdots$,

$$|g_{4n-1}(z)-f_{2n}(z)|=\frac{|d_{4n}z||f_{2n}(z)-f_{2n-1}(z)|}{|h_{2n}(z)-d_{4n}z|}, \qquad n\geq 1, \tag{4.55}$$

and

$$|g_{4n+1}(z)-f_{2n+1}(z)|=\frac{|d_{4n+2}z||f_{2n+1}(z)-f_{2n}(z)|}{|h_{2n+1}(z)-d_{4n+2}z|}, \qquad n\geq 1. \tag{4.56}$$

The partial numerators $a_n(z)$ of (4.54) converge to $z^2/4$ as $n\to\infty$. Therefore, by Theorem A, $\{h_n(z)\}$ converges to one of the two roots of

$$x^2-x-\frac{z^2}{4}=0$$

if $z\in\mathbb{C}-\{R[-i]\cup R[i]\}$. These roots are

$$\frac{1\pm\sqrt{1+z^2}}{2},$$

where $\sqrt{\ }$ denotes the square root with positive real part. In particular, if z is real, it is easy to see that $\lim_{n\to\infty} h_n(z)=(1+\sqrt{1+z^2})/2$ (since the other choice is a negative number). The sequences $\{d_{4n}z\}$ and $\{d_{4n+2}z\}$ converge to z/π and $\pi z/4$, respectively. We now see from (4.55) and (4.56) that the sequences $\{g_{4n-1}(z)\}$ and $\{g_{4n+1}(z)\}$ will converge to $\arctan z$ in any region $\{f_n(z)\}$ converges to this function, provided $h_{2n}(z)\nrightarrow z/\pi$ and $h_{2n+1}(z)\nrightarrow \pi z/4$ as $n\to\infty$. Thus we investigate the roots of the equations

$$\frac{1\pm\sqrt{1+z^2}}{2}=\frac{\pi z}{4} \quad \text{and} \quad \frac{1\pm\sqrt{1+z^2}}{2}=\frac{z}{\pi}.$$

The only possible candidates for roots of the first set of equations are

$$z=0 \quad \text{and} \quad z=\left(\frac{\pi}{4}-\frac{1}{\pi}\right)^{-1}\cong 2.140922923.$$

We have already disposed of the case $z=0$. For the second value of z above we obtain

$$\lim_{n\to\infty} h_n\left(\left(\frac{\pi}{4}-\frac{1}{\pi}\right)^{-1}\right)=\frac{1+\sqrt{1+(\pi/4-1/\pi)^{-2}}}{2}=\frac{\pi(\pi/4-1/\pi)^{-1}}{4},$$

so for $z=(\pi/4-1/\pi)^{-1}$ it follows that

$$\lim_{n\to\infty}(h_{2n}(z)-d_{4n+2}z)=0.$$

Therefore, unfortunately, our methods do not allow us to decide the convergence behavior of (4.50) for this value of z. The only possible roots of the second set of equations above having z/π on the right side are $z=0$ and $z=-(\pi/4-1/\pi)^{-1}$. But $\{h_{2n+1}(z)-d_{4n+2}(z)\}$ converges to a nonzero limit at $z=-(\pi/4-1/\pi)^{-1}$, so that the continued fraction (4.50) converges to $\arctan z$ at this point. Thus

$$\lim_{n\to\infty}g_n(z)=\lim_{n\to\infty}f_n(z)=\arctan z$$

over the region indicated in part (A) of our theorem.

The proof of part (B) will not be given since it parallels the proof of part (A). We give only the continued fraction representation of $\operatorname{arctanh} z$ from which the δ-fraction (4.52) can be derived by an extension. Making use of the fact that $\operatorname{arctanh} z=-i\arctan(iz)$ and employing (4.54), it follows that

$$\operatorname{arctanh} z=\frac{z}{1-}\ \frac{1^2z^2/1\times3}{1}\ -\ \frac{2^2z^2/3\times5}{1}\ -\ \frac{3^2z^2/5\times7}{1}\ -\ \frac{4^2z^2/7\times9}{1}\ -\cdots, \tag{4.57}$$

valid if $z\in\mathbb{C}-\{R[-1]\cup R[1]\}$.

The extension process used to derive the δ-fraction expansion for the function $I_3(z)$ in our next theorem is more complicated than in the previous examples. The expansion we give is nonregular. We were unable to determine a general formula for the coefficients of the regular δ-fraction expansion for this function. Our investigations seem to indicate that such a formula would be extremely complicated. The (6,6) limit periodic δ-fraction (4.58) in our next example is our first example of a δ-fraction expansion of an analytic function in which infinitely many of the δ_n's are not 0.

THEOREM 4.8. *If* $I_3(z)=\int_0^z dt/(1+t^3)$, *then a* δ*-fraction expansion of* $I_3(z)$ *is given by*

$$\begin{aligned}I_3(z)=&\frac{z}{1-z}+\frac{z}{1-z}+\frac{z}{1}+\frac{z/b_2}{1}\ -\ \frac{z/b_2}{1}+\frac{z}{1-z}-\frac{z/b_3}{1}\\&+\frac{z/b_3}{1-z}+\frac{z}{1}+\frac{z/b_4}{1}\ -\ \frac{z/b_4}{1}+\frac{z}{1-z}-\frac{z/b_5}{1}+\frac{z/b_5}{1-z}+\cdots,\end{aligned} \tag{4.58}$$

where for $n=1,2,\cdots$

(4.59)

$$b_1=1,\qquad b_{2n+1}=\frac{(6n+1)}{(3n+1)^2}\left[\prod_{k=1}^{n}\frac{3k+1}{3k}\right]^2=\frac{(6n+1)}{(3n+1)^2}\left[\frac{\Gamma(n+4/3)}{\Gamma(4/3)\Gamma(n+1)}\right]^2=O(n^{-1/3}),$$

$$b_2=4,\qquad b_{2n+2}=(6n+4)\left[\prod_{k=1}^{n}\frac{3k+1}{3k}\right]^2=(6n+4)\left[\frac{\Gamma(4/3)\Gamma(n+1)}{\Gamma(n+4/3)}\right]^2=O(n^{1/3}).$$

The representation (4.58) *is valid for all* z *such that* $z^3\in\mathbb{C}-R[-1]$, *or equivalently, for all* $z\in\mathbb{C}-\{R[-1]\cup R[\omega]\cup R[\omega^2]\}$, *where* ω *is a nonreal cube root of* -1. *The continued fraction* (4.58) *is a nonregular* (6,6) *limit periodic* δ*-fraction.*

Proof. According to [9, p. 203]

$$\begin{aligned}I_3(z)=&\frac{z}{1+}\ \frac{1^2z^3}{4}\ +\ \frac{3^2z^3}{7}\ +\ \frac{4^2z^3}{10}\ +\ \frac{6^2z^3}{13}\ +\ \frac{7^2z^3}{16}\\&+\cdots+\ \frac{(3n)^2z^3}{6n+1}\ +\ \frac{(3n+1)^2z^3}{6n+4}\ +\cdots,\end{aligned} \tag{4.60}$$

and this representation is valid for all z such that $z^3 \in \mathbb{C} - R[-1]$. With the aid of well-known equivalence transformations it can be shown that the continued fraction (4.60) is equivalent to

$$I_3(z) = \frac{z}{1} + \frac{1^2z^3/4}{1} + \frac{3^2z^3/4\times 7}{1} + \frac{4^2z^3/7\times 10}{1} + \frac{6z^3/10\times 13}{1} + \frac{7^2z^3/13\times 16}{1} + \cdots + \frac{(3n)^2z^3/(6n-2)(6n+1)}{1} + \frac{(3n+1)^2z^3/(6n+1)(6n+4)}{1} + \cdots, \tag{4.61}$$

which in turn is equivalent to

$$I_3(z) = \frac{z}{b_1} + \frac{z^3}{b_2} + \frac{z^3}{b_3} + \frac{z^3}{b_4} + \cdots, \tag{4.62}$$

where the b_n are given by (4.59). We now use a method of repeated extensions to derive the continued fraction (4.58) from the continued fraction (4.62). First, we extend (4.62) to obtain

$$\frac{z}{1-z} + \frac{z}{1} - \frac{z^2}{b_2} + \frac{z^2}{1} - \frac{z}{b_3} + \frac{z}{1} - \frac{z^2}{b_4} + \frac{z^2}{1} - \frac{z}{b_5} + \frac{z}{1} - \cdots. \tag{4.63}$$

Let $A_n(z)$ ($B_n(z)$) denote the nth numerator (denominator) of (4.62). Then the approximants of (4.63) are given by

$$\frac{A_1 - zA_0}{B_1 - zB_0}, \frac{A_1}{B_1}, \frac{A_2 - z^2A_1}{B_2 - z^2B_1}, \frac{A_2}{B_2}, \frac{A_3 - zA_2}{B_3 - zB_2}, \frac{A_3}{B_3}, \frac{A_4 - z^2A_3}{B_4 - z^2B_3}, \frac{A_4}{B_4}, \ldots.$$

We relabel these approximants as

$$\frac{P_1}{Q_1}, \frac{P_2}{Q_2}, \frac{P_3}{Q_3}, \frac{P_4}{Q_4}, \frac{P_5}{Q_5}, \frac{P_6}{Q_6}, \frac{P_7}{Q_7}, \frac{P_8}{Q_8}, \ldots,$$

respectively, and we assume that corresponding numerators and denominators are equal. Next, we extend (4.63) to obtain

$$\frac{z}{1-z} + \frac{z}{1-z} + \frac{z}{1} + \frac{z}{b_2 - z} + \frac{z^2}{1} - \frac{z}{b_3} + \frac{z}{1-z} + \frac{z}{1} + \frac{z}{b_4 - z} + \frac{z^2}{1} - \frac{z}{b_5} + \frac{z}{1-z} + \cdots, \tag{4.64}$$

whose approximants are

$$\frac{P_1}{Q_1}, \frac{P_2 - zP_1}{Q_2 - zP_1}, \frac{P_2}{Q_2}, \frac{P_3}{Q_3}, \frac{P_4}{Q_4}, \frac{P_5}{Q_5}, \frac{P_6 - zP_5}{Q_6 - zQ_5}, \frac{P_6}{Q_6}, \frac{P_7}{Q_7}, \frac{P_8}{Q_8}, \frac{P_9}{Q_9}, \frac{P_{10} - zP_9}{Q_{10} - zQ_9}, \ldots.$$

We relabel these approximants by

$$\frac{N_1}{D_1}, \frac{N_2}{D_2}, \frac{N_3}{D_3}, \frac{N_4}{D_4}, \frac{N_5}{N_5}, \ldots,$$

respectively, as we did in our first extension above. Finally, we extend (4.64) to obtain

$$\frac{z}{1-z} + \frac{z}{1-z} + \frac{z}{1} + \frac{z}{b_2} - \frac{z}{1} + \frac{z}{1-z} - \frac{z}{b_3} + \frac{z}{1-z} + \frac{z}{1} + \frac{z}{b_4} - \frac{z}{1} + \frac{z}{1-z} - \frac{z}{b_5} + \cdots, \tag{4.65}$$

whose approximants are

$$\frac{N_1}{D_1}, \frac{N_2}{D_2}, \frac{N_3}{D_3}, \frac{N_4+zN_3}{D_4+zD_3}, \frac{N_4}{D_4}, \frac{N_5}{D_5}, \frac{N_6}{D_6}, \frac{N_7}{D_7}, \frac{N_8}{D_8}, \frac{N_9+zN_8}{D_9+zD_8}, \frac{N_9}{D_9}, \frac{N_{10}}{D_{10}}, \frac{N_{11}}{D_{11}}, \ldots.$$

The continued fraction (4.58) is obtained from (4.65) by an equivalence transformation. Let $g_n(z)$ denote the nth approximant of the continued fraction (4.58). Then, through keeping track of the relations between approximants in the extension process above, we are able to say that

(4.66)

$$g_{6n}(z)=\frac{A_{2n}(z)}{B_{2n}(z)}, \qquad g_{6n-1}(z)=\frac{A_{2n}(z)-z^2A_{2n-1}(z)}{B_{2n}(z)-z^2B_{2n-1}(z)},$$

$$g_{6n-2}=\frac{A_{2n}(z)+z(1-z)A_{2n-1}(z)}{B_{2n}(z)+z(1-z)B_{2n-1}(z)}, \qquad g_{6n+1}(z)=\frac{A_{2n+1}(z)-zA_{2n}(z)}{B_{2n+1}(z)-zB_{2n}(z)},$$

$$g_{6n+2}(z)=\frac{(1-z)A_{2n+1}(z)+z^2A_{2n}(z)}{(1-z)B_{2n+1}(z)+z^2B_{2n}(z)}, \qquad g_{6n+3}(z)=\frac{A_{2n-1}(z)}{B_{2n-1}(z)}.$$

Hence, in particular, $g_{3n}(z)=A_n(z)/B_n(z)\to I_3(z)$ as $n\to\infty$ if $z^3\in\mathbb{C}-R[-1]$. Let $f_n(z)=A_n(z)/B_n(z)$ and let $h_n(z)=B_n^*(z)/B_{n-1}^*(z)$, where $B_n^*(z)$ is the nth denominator of (4.61). Then, with the aid of (4.66) and the fundamental formulas for continued fractions, we obtain

$$|g_{6n+1}(z)-f_{2n+1}(z)|=\frac{|z||f_{2n+1}(z)-f_{2n}(z)|}{|b_{2n+1}h_{2n+1}(z)-z|},$$

$$|g_{6n+2}(z)-f_{2n+1}(z)|=\frac{|z^2/(1-z)||f_{2n+1}(z)-f_{2n}(z)|}{|b_{2n+1}h_{2n+1}(z)+z^2/(1-z)|},$$

$$|g_{6n-1}(z)-f_{2n}(z)|=\frac{|z^2/b_{2n}||f_{2n}(z)-f_{2n-1}(z)|}{|h_{2n}(z)-z^2/b_{2n}|},$$

$$|g_{6n-2}(z)-f_{2n}(z)|=\frac{|z(1-z)/b_{2n}||f_{2n}(z)-f_{2n-1}(z)|}{|h_{2n}(z)+z(1-z)/b_{2n}|}.$$

The continued fraction (4.61) is limit periodic because its partial numerators converge to the limit $z^3/4$ as $n\to\infty$. Therefore, by Theorem A, the sequence $\{h_n(z)\}$ converges to a root of the equation

$$x^2-z-\frac{z^3}{4}=0,$$

whose roots are

$$\frac{1\pm\sqrt{1+z^3}}{2},$$

if $\sqrt{\ }$ denotes the square root of $(1+z^3)$ with positive real part. Thus $\lim_{n\to\infty}h_n(z)=0$ only if $z=0$ when z is any complex number satisfying $z^3\in\mathbb{C}-R[-1]$. Clearly, (4.58) converges to 0 at $z=0$ as desired, so assume now that $z\neq 0$ and $z^3\in\mathbb{C}-R[-1]$. Then, since $b_{2n+1}\to 0$ and $b_{2n}\to\infty$ as $n\to\infty$, it follows from the last four formulas above and

the convergence behavior of $\{h_n(z)\}$ that the sequences $\{g_n(z)\}$ and $\{f_n(z)\}$ converge to the same limit $I_3(z)$. This completes our argument.

In attempting to find a regular δ-fraction expansion for the function

$$I_4(z)=\int_0^z \frac{dt}{1+t^4},$$

which is a generalization of $\arctan z$ as is $I_3(z)$, we ran into the same difficulties as we did for $I_3(z)$. However, we were able to find a regular δ-fraction representation for $z^2 I_4(z)$. This expansion, which surprisingly to us turns out to be (8, 1) limit periodic, is given in our next theorem. This example also leads us to believe that, for certain classes of functions, regular δ-fraction representations can be found for practical variations of these functions, if not for the functions themselves.

THEOREM 4.9. *Let $I_4(z)=\int_0^z dt/(1+t^4)$ and let $E_4(z)=z^2 I_4(z)$. Then the regular δ-fraction expansion of $E_4(z)$ is given by*

$$\begin{aligned}(4.67)\qquad E_4(z)=&-z+\frac{z}{1-z}+\frac{z}{1}+\frac{z/b_1}{1}-\frac{z/b_1}{1}+\frac{z}{1}-\frac{z}{1}-\frac{z/b_2}{1}\\&+\frac{z/b_2}{1}-\frac{z}{1}+\frac{z}{1}+\frac{z/b_3}{1}-\frac{z/b_3}{1}+\frac{z}{1}-\frac{z}{1}-\frac{z/b_4}{1}+\frac{z/b_4}{1}\\&-\frac{z}{1}+\frac{z}{1}+\frac{z/b_5}{1}-\frac{z/b_5}{1}+\frac{z}{1}-\frac{z}{1}-\frac{z/b_6}{1}+\frac{z/b_6}{1}-\cdots,\end{aligned}$$

where for $n=1,2,\cdots$

(4.68)

$$b_1=1,\qquad b_{2n+1}=\frac{(8n+1)}{(4n+1)^2}\left[\prod_{k=1}^{n}\frac{4k+1}{4k}\right]^2=\frac{(8n+1)}{(4n+1)^2}\left[\frac{\Gamma(n+5/4)}{\Gamma(5/4)\Gamma(n+1)}\right]^2=O(n^{-1/2}),$$

$$b_2=5,\qquad b_{2n+2}=(8n+5)\left[\prod_{k=1}^{n}\frac{4k}{4k+1}\right]^2=(8n+5)\left[\frac{\Gamma(5/4)\Gamma(n+1)}{\Gamma(n+5/4)}\right]^2=O(n^{1/2}).$$

The expansion (4.67) *is valid for all z such that $z^4\in\mathbb{C}-R[-1]$, i.e., for all z such that $z\in\mathbb{C}-\bigcup_{k=1}^{4}R[\omega^k]$, where $\omega=(1+i)/\sqrt{2}$. The δ-fraction* (4.67) *is* (8, 1) *limit periodic.*

Proof. Using [9, formula 6.1.19, p. 203] we obtain

$$\begin{aligned}(4.69)\qquad E_4(z)=&\frac{z^3}{1}+\frac{z^4}{5}+\frac{4^2z^4}{9}+\frac{5^2z^4}{13}\\&+\frac{8^2z^4}{17}+\frac{9^2z^4}{21}+\cdots+\frac{(4n)^2z^4}{8n+1}+\frac{(4n+1)^2z^4}{8n+5}+\cdots,\end{aligned}$$

and this representation is valid for all z such that $z^4\in\mathbb{C}-R[-1]$. The continued fraction (4.68) is equivalent to each of the continued fractions

(4.70)

$$E_4(z)=\frac{z^3}{1}+\frac{z^4/5}{1}+\frac{4^2z^4/5\times 9}{1}+\frac{5^2z^4/9\times 13}{1}+\frac{8^2z^4/13\times 17}{1}$$

$$+\frac{9^2z^4/17\times 21}{1}+\cdots+\frac{(4n)^2z^4/(8n-3)(8n+1)}{1}+\frac{(4n+1)^2z^4/(8n+1)(8n+5)}{1}+\cdots$$

and

$$(4.71)\qquad E_4(z)=\frac{z^3}{b_1}+\frac{z^4}{b_2}+\frac{z^4}{b_3}+\cdots,$$

where the b_n are given by (4.68). As in the proof of Theorem 4.8, we use the method of repeated extensions to arrive at the δ-fraction (4.67). We extend (4.71) by choosing $\rho_0=z$, $\rho_n=z^2$ $(n\geq 1)$ in Theorem B to obtain

$$-z+\frac{z}{1-b_1}+\frac{z^2}{1}-\frac{z^2}{b_2}+\frac{z^2}{1}-\frac{z^2}{b_3}+\frac{z^2}{1}-\frac{z^2}{b_4}+\cdots. \tag{4.72}$$

We then extend (4.72) and arrive at

$$\begin{aligned}-z+\frac{z}{1-z}+\frac{z}{1}+\frac{z}{b_1}-\frac{z}{1}+\frac{z}{1}-\frac{z}{1}-\frac{z}{b_2}+\frac{z}{1}\\ -\frac{z}{1}+\frac{z}{1}+\frac{z}{b_3}-\frac{z}{1}+\frac{z}{1}-\frac{z}{1}-\frac{z}{b_4}+\frac{z}{1}-\frac{z}{1}+\frac{z}{1}\\ +\frac{z}{b_5}-\frac{z}{1}-\frac{z}{1}-\frac{z}{b_6}+\cdots.\end{aligned} \tag{4.73}$$

The continued fraction (4.67) is derived from (4.73) by the appropriate equivalence transformation. Let $A_n(z)(B_n(z))$ denote the nth numerator (denominator) of (4.71), and let $g_n(z)$ denote the nth approximant of (4.67). Then

$$\begin{aligned}
&g_{8n-1}(z)=\frac{A_{2n}-z(1+z)A_{2n-1}}{B_{2n}-z(1+z)B_{2n-1}}, && g_{8n+3}(z)=\frac{A_{2n+1}+z(1-z)A_{2n}}{B_{2n+1}+z(1-z)B_{2n}},\\
&g_{8n}(z)=\frac{A_{2n}-z^2A_{2n-1}}{B_{2n}-z^2B_{2n-1}}, && g_{8n+4}(z)=\frac{A_{2n+1}-z^2A_{2n}}{B_{2n+1}-z^2B_{2n}},\\
&g_{8n+1}(z)=\frac{A_{2n}(1-z)+z^3A_{2n-1}}{B_{2n}(1-z)+z^3B_{2n-1}} && g_{8n+5}(z)=\frac{A_{2n+1}(1+z)-z^3A_{2n}}{B_{2n+1}(1+z)-z^3B_{2n}},\\
&g_{8n+2}(z)=\frac{A_{2n}}{B_{2n}}, && g_{8n+6}(z)=\frac{A_{2n+1}}{B_{2n+1}}.
\end{aligned} \tag{4.74}$$

Now let $f_n(z)=A_n(z)/B_n(z)$ and $h_n(z)=B_n^*(z)/B_{n-1}^*(z)$, where $B_n^*(z)$ is the nth denominator of (4.70). Then

$$b_nh_n(z)=\frac{B_n(z)}{B_{n-1}(z)}.$$

It follows that, if $m=-1,0$, or 1, then

$$|g_{8n+m}(z)-f_{2n}(z)|=\frac{|t_m(z)/b_{2n}||f_{2n}(z)-f_{2n-1}(z)|}{|h_{2n}(z)-t_m(z)/b_{2n}|} \tag{4.75}$$

where

$$t_{-1}(z)=z(1+z),\quad t_0(z)=z^2,\quad t_1(z)=-\frac{z^3}{1-z},$$

and, if $m=3,4$ or 5, then

$$|g_{8n+m}(z)-f_{2n+1}(z)|=\frac{|t_m(z)||f_{2n+1}(z)-f_{2n}(z)|}{|b_{2n+1}h_{2n+1}(z)-t_m(z)|}, \tag{4.76}$$

where

$$t_3(z)=-z(1-z),\quad t_4(z)=z^2,\quad t_5(z)=\frac{z^3}{1+z}.$$

The partial numerators of (4.70) converge to $z^4/4$ as $n\to\infty$, so for all z such that $z^4\in\mathbb{C}-R[-1]$ it follows from Theorem A that $\{h_n(z)\}$ must converge to one of the two numbers $(1\pm\sqrt{1+z^4})/2$. We note that neither one of these numbers is 0 unless $z=0$, but for this value of z, (4.67) is clearly valid. It follows from (4.74), (4.75), (4.76) and the limit behavior of the b_n ($b_{2n}\to\infty$ and $b_{2n+1}\to 0$ as $n\to\infty$) that

$$\lim_{n\to\infty} g_{8n+m}(z)=\lim_{n\to\infty} f_n(z)=E_4(z),\qquad m=-1,0,\cdots,6.$$

Hence, our representation (4.67) is valid as asserted.

In our next theorem we give a new continued fraction representation for the function $\exp(z^2)$ that turns out to be a regular δ-fraction with all of its partial denominators equal to 1. The proof that the corresponding δ-fraction for $\exp(z^2)$ converges to $\exp(z^2)$ everywhere demands a new twist not employed in our previous examples.

THEOREM 4.10. (A) *The regular δ-fraction expansion of* $\exp(z)$, *valid for all* $z\in\mathbb{C}$, *is given by*

$$\exp(z)=1+\frac{z}{1}-\frac{z/2}{1}+\frac{z/6}{1}-\frac{z/6}{1}+\frac{z/10}{1}-\frac{z/10}{1}+\frac{z/14}{1}-\frac{z/14}{1}+\cdots. \tag{4.77}$$

(B) *The regular δ-fraction expansion of* $\exp(z^2)$, *valid for all* $z\in\mathbb{C}$ *is given by*

$$\begin{aligned}\exp(z^2)=(1-z)&+\frac{z}{1}-\frac{z/1}{1}+\frac{z/1}{1}+\frac{z/2}{1}-\frac{z/2}{1}+\frac{z/3}{1}-\frac{z/3}{1}\\&-\frac{z/2}{1}+\frac{z/2}{1}-\frac{z/5}{1}+\frac{z/5}{1}+\frac{z/2}{1}-\frac{z/2}{1}+\frac{z/7}{1}-\frac{z/7}{1}\\&-\frac{z/2}{1}+\frac{z/2}{1}-\frac{z/9}{1}+\frac{z/9}{1}+\frac{z/2}{1}-\frac{z/2}{1}+\cdots\\&=(1-z)+\mathop{\mathrm{K}}_{n=1}^{\infty}\left(\frac{d_n z}{1}\right),\end{aligned} \tag{4.78}$$

where for $n=1,2,\cdots$,

$$\begin{aligned}&d_1=1, && d_{8n-4}=\frac{1}{2}, && d_{8n-1}=-\frac{1}{4n-1},\\&d_{8n-6}=-\frac{1}{4n-3}, && d_{8n-3}=-\frac{1}{2}, && d_{8n}=-\frac{1}{2},\\&d_{8n-5}=\frac{1}{4n-3}, && d_{8n-2}=\frac{1}{4n-1}, && d_{8n+1}=\frac{1}{2}.\end{aligned} \tag{4.79}$$

The δ-fraction (4.78) *is* (8, 1) *limit periodic.*

Proof. It is well known that the following representation for $\exp(z)$ is valid for all $z\in\mathbb{C}$:

$$\exp(z)=1+\frac{z}{1}-\frac{z}{2}+\frac{z}{3}-\frac{z}{2}+\frac{z}{5}-\cdots-\frac{z}{2}+\frac{z}{2n+1}-\cdots. \tag{4.80}$$

Two sources giving this expansion are [9, p. 207] and [12, p. 113]. The representation (4.77) is easily derived from (4.80) by an equivalence transformation, which may be found in [21, p. 124]. By replacing z by z^2 in (4.80) and using another equivalence

transformation we have for all $z \in \mathbb{C}$ that

$$\exp(z^2) = 1 + \frac{z^2}{1} + \frac{z^2}{(-2)} + \frac{z^2}{(-3)} + \frac{z^2}{2} + \frac{z^2}{5} + \frac{z^2}{(-2)} + \frac{z^2}{(-7)} + \cdots \tag{4.81}$$

$$= 1 + \mathop{\mathrm{K}}_{n=1}^{\infty} \left(\frac{z^2}{b_n} \right),$$

where

$$b_{2n} = (-1)^n 2 \quad \text{and} \quad b_{2n-1} = (-1)^{n-1}(2n-1), \qquad n \geq 1. \tag{4.82}$$

We extend the continued fraction (4.81) (using Theorem B with $\rho_n \equiv z$) to obtain the continued fraction

$$(1-z) + \mathop{\mathrm{K}}_{n=1}^{\infty} \left(\frac{(-1)^{n-1} z}{c_n} \right), \quad \text{where } c_{2n-1} \equiv 1 \quad \text{and} \quad c_{2n} \equiv b_n. \tag{4.83}$$

The regular δ-fraction

$$(1-z) + \frac{z}{1} - \frac{z/b_1}{1} + \frac{z/b_1}{1} - \frac{z/b_2}{1} + \frac{z/b_2}{1} - \frac{z/b_3}{1} + \frac{z/b_3}{1} - \cdots, \tag{4.84}$$

where the b_n are given by (4.82), is equivalent to (4.83). It is easily verified that the continued fractions (4.84) and (4.78) are identical. Let $g_n(z)$ denote the nth approximant of (4.78) and let $A_n(z)(B_n(z))$ denote the nth numerator (denominator) of

$$\exp(z^2) = 1 + \frac{z^2}{1} - \frac{z^2/2}{1} + \frac{z^2/6}{1} - \frac{z^2/6}{1} + \frac{z^2/10}{1} - \frac{z^2/10}{1} + \cdots \tag{4.85}$$

derived from (4.77) by replacing z by z^2.
Then (if $f_n(z) = A_n(z)/B_n(z)$ and $h_n(z) = B_n(z)/B_{n-1}(z)$),

$$g_{2n+1}(z) = f_n(z) \quad (n \geq 0), \quad \text{and} \quad g_{2n}(z) = \frac{A_n(z) - (z/b_n)A_{n-1}(z)}{B_n(z) - (z/b_n)B_{n-1}(z)} \qquad (n \geq 1).$$

Hence,

$$|g_{8n}(z) - f_{4n}(z)| = \frac{|z/2||f_{4n}(z) - f_{4n-1}(z)|}{|h_{4n}(z) - z/2|}, \tag{4.86}$$

$$|g_{8n+2}(z) - f_{4n+1}(z)| = \frac{|z/(4n+1)||f_{4n+1}(z) - f_{4n}(z)|}{|h_{4n+1}(z) - z/(4n+1)|}, \tag{4.87}$$

$$|g_{8n+4}(z) - f_{4n+2}(z)| = \frac{|z/2||f_{4n+2}(z) - f_{4n+1}(z)|}{|h_{4n+2}(z) + z/2|}, \tag{4.88}$$

$$|g_{8n+6}(z) - f_{4n+3}(z)| = \frac{|z/(4n+3)||f_{4n+3}(z) - f_{4n+2}(z)|}{|h_{4n+3}(z) + z/(4n+3)|}. \tag{4.89}$$

The sequence of partial numerators of (4.85) converges to 0 as $n \to \infty$. Therefore, it follows from Theorem A that $\lim_{n\to\infty} h_n(z) = 0$ or 1. Unfortunately, for arbitrary z, we do not know how to determine whether $\{h_n(z)\}$ converges to 0 or 1. Hence, we cannot employ expressions (4.87) and (4.89) above and the technique we used in the past to determine the convergence behavior of $\{g_{8n+2}(z)\}$ and $\{g_{8n+6}(z)\}$. Therefore, we use a

different approach here and investigate first the convergence of a tail of (4.78). Let

$$\mathsf{K}=\underset{m=1}{\overset{\infty}{\mathsf{K}}}\left(\frac{zd_{8n-4+m}}{1}\right),$$

where the coefficients of z are given by (4.79). Then the odd part K_n^* of K_n is given by

$$\mathsf{K}_n^*=-\frac{z}{2}+\frac{z^2/2(4n-1)}{1}-\frac{z^2/2(4n-1)}{1}+\frac{z^2/2(4n+1)}{1}-\frac{z^2/2(4n+1)}{1}$$
$$+\frac{z^2/2(4n+3)}{1}-\frac{z^2/2(4n+3)}{1}+\frac{z^2/2(4n+5)}{1}-\frac{z^2/2(4n+5)}{1}+\cdots.$$

Suppose $|z|\leq M(>0)$ and choose n in K_n^* large enough such that

$$\frac{M^2}{2(4n-1)}\leq\frac{1}{4}.$$

It follows from the convergence neighborhood theorem (see [9 p. 108]) and Theorem 3.2 that K_n^* converges uniformly on the set $|z|\leq M$ to a function $F_n(z)$ that is analytic on $|z|<M$. Now let G_m denote the mth approximant of K_n, and let $N_m(z)(D_m(z))$ denote the nth numerator (denominator) of K_n^*. Then

$$G_{2m+1}(z)=\frac{N_m(z)}{D_m(z)}\quad\text{and}\quad G_{2m}(z)=\frac{N_m(z)-zd_{8n-3+2m}N_{m-1}(z)}{D_m(z)-zd_{8n-3+2m}D_{m-1}(z)},$$

with the aid of which we obtain

$$\left|G_{2m}(z)-\frac{N_m(z)}{D_m(z)}\right|=\frac{|zd_{8n-3+2m}||N_m(z)/D_m(z)-N_{m-1}(z)/D_{m-1}(z)|}{|(D_m(z)/D_{m-1}(z))-zd_{8n-3+2m}|}.\tag{4.90}$$

If we let $H_m(z)=D_m(z)/D_{m-1}(z)$, then $H_1(z)\equiv 1$ and it can be verified by induction that

$$|H_{2m+1}(z)-1|\leq\frac{4n-1}{2[4n+2m-1]}\quad\text{and}\quad|H_{2m+2}(z)-1|\leq\frac{4n-1}{2[4n+2m-1]},$$

provided z satisfies $|z|\leq M$. Hence, we can now say that $\lim_{m\to\infty}H_m(z)=1$. By splitting $\{G_{2m}(z)\}$ into four subsequences as we did for $\{g_{2m}(z)\}$ and by using (4.90), it can be seen without too much difficulty that $\lim_{m\to\infty}G_{2m}(z)=\lim_{m\to\infty}N_m(z)/D_m(z)=F_n(z)$ if $|z|<M$. Thus given an arbitrary positive number M there is a tail of (4.78) which converges to an analytic function F_n on $|z|<M$. By the convergence neighborhood theorem [9, p. 108] applied to (4.78), we have, in particular, that (4.78) converges uniformly on the set $|z|\leq\frac{1}{2}$. Hence, by arguments similar to those used in the proof of Theorem 3.5, the δ-fraction (4.78) must converge to $\exp(z^2)$ in a neighborhood of the origin and, therefore, to $\exp(z^2)$ throughout $|z|<M$. Since M was arbitrary, our proof is complete.

We now give a brief discussion of some connections between certain associated continued fractions and δ-fractions. An associated continued fraction is a continued fraction of the form

$$1+\frac{k_1z}{1+l_1z}-\frac{k_2z^2}{1+l_2z}-\frac{k_3z^2}{1+l_3z}-\frac{k_4z^2}{1+l_4z}-\cdots,\tag{4.91}$$

where $k_\nu\neq 0$, $k=1,2,\cdots$. Jones and Thron [9, §7.2] and Perron [21, §25] give thorough treatments of these continued fractions in their books. A number of the δ-fraction

expansions of analytic functions already given in this section were derived from associated continued fraction representations of these functions of the type where $l_1 \equiv 0$. So, we concern ourselves here with the case where $l_i \neq 0$, $i = 1, 2, \cdots$. Then by extending (4.91) we obtain

(4.92)

$$1 + \cfrac{k_1 z}{1} + \cfrac{l_1 z}{1} + \cfrac{k_2 z / l_1}{1} + \cfrac{(l_2 - k_2/l_1)z}{1} + \cfrac{(k_3/(l_2 - k_2/l_1))z}{1} + \cfrac{(l_3 - k_3/(l_2 - k_2/l_1))z}{1} + \cfrac{(k_4/(l_3 - k_3/(l_2 - k_2/l_1)))z}{1} + \cfrac{(l_4 - k_4/(l_3 - k_3/(l_2 - k_2/l_1)))z}{1} + \cdots,$$

provided also that none of the coefficients of z in the partial numerators of (4.92) are zero. It is interesting that these coefficients are finite continued fractions themselves. The continued fractions (4.91) and (4.92) are equivalent in the sense that they correspond to the same power series at $z = 0$. If we let $\alpha = 0$, $\beta = 1$, $a = 1 + c$, $b = 1 - c$, where $0 < c < 1$, in the example given by Perron [21, p. 261], and if

$$L_c(z) = 1 + \log\left[\frac{1 + (1+c)z}{1 + (1-c)z}\right], \tag{4.93}$$

we arrive at

$$L_c(z) \sim 1 + \cfrac{k_1 z}{1+z} - \cfrac{k_2 z^2}{1+z} - \cfrac{k_3 z^2}{1+z} - \cdots, \tag{4.94}$$

where

$$k_1 = 2c \quad \text{and} \quad k_\nu = \frac{(\nu-1)^2 c^2}{(2\nu-1)(2\nu-3)}, \qquad \nu \geq 2. \tag{4.95}$$

Using the techniques of this section and the convergence results of the preceding section, particularly Poincaré's theorem, we can say the following:

THEOREM 4.11. *The regular δ-fraction expansion of the function $L_c(z)$ defined by* (4.93) *is given by*

(4.96)

$$L_c(z) = 1 + \cfrac{k_1 z}{1} + \cfrac{z}{1} + \cfrac{k_2 z}{1} + \cfrac{(1-k_2)z}{1} + \cfrac{(k_3/(1-k_2/1))z}{1} + \cfrac{(1-k_3/(1-k_2/1))z}{1} + \cfrac{(k_4/(1-k_3/(1-k_2/1)))z}{1} + \cfrac{(1-k_4/(1-k_3/(1-k_2/1)))z}{1} + \cdots,$$

where the k_ν are given by (4.95). *The representation* (4.96) *is valid for all $z \in \mathbb{C} - [-1/(1-c), 1/(1+c)]$, where c satisfies $0 < c < 1$.*

In our next theorem we give two examples of a regular δ-fraction corresponding to a divergent asymptotic series connected with the Laplace transform of a certain function. As we shall see, though the series diverges, the corresponding δ-fraction converges to a function that can be used in the evaluation of the Laplace transform. Before we give these examples, let us recall that the Bernoulli numbers B_n are defined by

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} \left(\frac{B_n}{n!}\right) x^n, \tag{4.97}$$

giving, in particular, $B_0=1$, $B_1=-\frac{1}{2}$, $B_{2k+1}=0$ $(k\geq 1)$, $B_2=\frac{1}{6}$, $B_4=-\frac{1}{30}$, $B_6=\frac{1}{42}$, and $B_8=-\frac{1}{30}$.

THEOREM 4.12. (A) *If*

$$\mathsf{K}(x)=-x+\frac{x}{1}-\frac{x/(1+\frac{1}{2})}{1}+\frac{x/(1+\frac{1}{2})}{1}-\frac{x/(\frac{1}{2}+\frac{1}{3})}{1}+\frac{x/(\frac{1}{2}+\frac{1}{3})}{1}$$
$$-\frac{x/(\frac{1}{3}+\frac{1}{4})}{1}+\frac{x/(\frac{1}{3}+\frac{1}{4})}{1}-\cdots,$$

then $\mathsf{K}(x)$ *converges for all* $x\leq 0$,

$$\mathsf{K}(x)\sim\sum_{k=1}^{\infty}B_{2k}(2x)^{2k},$$

and

$$\frac{1}{s}\mathsf{K}\left(-\frac{1}{s}\right)=\int_0^{\infty}e^{-st}(-1+t\coth t)\,dt\quad(s>0),$$

where B_{2k} *is the* $2k$*th Bernoulli number defined by* (4.97).

(B) *If*

$$\mathsf{K}^*(x)=-x+\frac{x}{1}-\frac{x}{1}+\frac{x}{1}-\frac{2x}{1}+\frac{2x}{1}-\frac{3x}{1}+\frac{3x}{1}$$
$$-\frac{4x}{1}+\frac{4x}{1}-\frac{5x}{1}+\frac{5x}{1}-\cdots,$$

then $\mathsf{K}^*(x)$ *converges for all* $x\leq 0$,

$$\mathsf{K}^*(x)\sim\sum_{k=1}^{\infty}\frac{B_{2k}}{2k}(2^{2k}-1)(2x)^{2k},$$

and

$$\mathsf{K}^*\left(-\frac{1}{s}\right)=\int_0^{\infty}e^{-st}\tanh t\,dt\qquad(s>0),$$

where again B_{2k} *is the* $2k$*th Bernoulli number.*

Proof. Recently, Frame [4] proved that the continued fraction

$$B(x)=\frac{x^2}{\frac{1}{1}+\frac{1}{2}}+\frac{x^2}{\frac{1}{2}+\frac{1}{3}}+\frac{x^2}{\frac{1}{3}+\frac{1}{4}}+\cdots$$

corresponds to the divergent power series $\sum_{k=1}^{\infty}B_{2k}(2x)^{2k}$, where the B_{2k} are Bernoulli numbers. The following relation is also given in [4]:

$$\frac{1}{s}B\left(\frac{1}{s}\right)=\int_0^{\infty}e^{-st}[1-t\coth t]\,dt\quad(s>0).\tag{4.98}$$

We extend the continued fraction $B(x)$ to obtain

$$-x+\frac{x}{1}-\frac{x}{\frac{1}{1}+\frac{1}{2}}+\frac{x}{1}-\frac{x}{\frac{1}{2}+\frac{1}{3}}+\frac{x}{1}-\frac{x}{\frac{1}{3}+\frac{1}{4}}+\frac{x}{1}-\cdots.\tag{4.99}$$

The continued fraction $\mathsf{K}(x)$ in part (A) is obtained from (4.99) by an equivalence transformation. It will be convenient for us now to write $B(x)$ in the equivalent form

$$B(x)=\frac{x^2/(1+\frac{1}{2})}{1}+\frac{x^2/(1+\frac{1}{2})(\frac{1}{2}+\frac{1}{3})}{1}+\frac{x^2/(\frac{1}{2}+\frac{1}{3})(\frac{1}{3}+\frac{1}{4})}{1}+\cdots.\tag{4.100}$$

If $f_n(x)$ and $g_n(x)$ denote the nth approximants of (4.100) and $\mathsf{K}(x)$ respectively, and if $D_n(x)$ denotes the denominator of $f_n(x)$, then

(4.101)

$$g_{2n+1}(x)=f_n(x) \quad \text{and} \quad |g_{2n}(x)-f_n(x)|=\frac{|x||f_n(x)-f_{n-1}(x)|}{|(1/n+1/(n+1))(D_n(x)/D_{n-1}(x))-x|}, \qquad n\geq 1.$$

Since $D_n(x)/D_{n-1}(x)\geq 1$ for all real x, it follows from (4.101) that

$$|g_{2n}(x)-f_n(x)|\leq|f_n(x)-f_{n-1}(x)|,$$

provided $x<0$. Hence from this inequality and (4.101), we have that

$$\lim_{n\to\infty} g_n(x)=\lim_{n\to\infty} f_n(x)=B(x)$$

for all $x<0$. The fact that $(1/s)\mathsf{K}(-1/s)$ is the Laplace transform of $-1+t\coth t$ now follows immediately from (4.98), after we note that $B(x)=B(-x)$. This completes our argument for part (A). The proof of part (B) is similar to the proof for (A), so we give only a sketch of how $\mathsf{K}^*(x)$ and it corresponding series can be derived. An equivalent form of the following representation can be found in Wall [33, p. 369].

(4.102)
$$\int_0^\infty e^{-zu}\tanh u\,du=\frac{z^{-2}}{1}+\frac{z^{-2}}{\frac{1}{2}}+\frac{z^{-2}}{\frac{1}{3}}+\frac{z^{-2}}{\frac{1}{4}}+\cdots.$$

By expanding $e^{-zu}\tanh u$ into a power series in u and then integrating from 0 to ∞ term by term one gets the series

(4.103)
$$\sum_{k=1}^{\infty}\frac{B_{2k}}{2k}(2^{2k}-1)2^{2k}z^{-2k}.$$

After setting $z=1/x$ in (4.102) and (4.103) we arrive at

$$\sum_{k=1}^{\infty}\frac{B_{2k}}{2k}(2^{2k}-1)(2x)^{2k}\sim\widehat{\mathsf{K}}(x),$$

where

$$\widehat{\mathsf{K}}(x)=\frac{x^2}{1}+\frac{x^2}{\frac{1}{2}}+\frac{x^2}{\frac{1}{3}}+\cdots+\frac{x^2}{\frac{1}{n}}+\cdots.$$

The continued fraction $\mathsf{K}^*(x)$ is derived from $\widehat{\mathsf{K}}(x)$.

To give some indication of how complicated the formulas can be for the coefficients of the regular δ-fraction corresponding to the Maclaurin series of a much used analytic function, we close this section with an informal and somewhat incomplete discussion of what happens in the case of $\sin z$. Recently, Dzjadyk [3] gave the associated continued fraction representation

(4.104)
$$\sin z=\frac{z}{1}+\frac{d_1z^2}{1}+\frac{d_2z^2}{1}+\frac{d_3z^2}{1}+\frac{d_4z^2}{1}+\cdots,$$

where

$$d_1=\frac{1}{6},\quad d_2=-\frac{7}{60},\quad d_3=\frac{11}{980},\quad d_4=-\frac{551}{19404},$$

and where the general formulas given for the d_n are so complicated that we choose not to repeat them here. Another important claim made in [3] is that none of the d_n's is 0.

Before we proceed further, let us introduce the following sequences $\{\phi_\nu\}$ and $\{\psi_\nu\}$ of Hankel determinants: Given an arbitrary sequence $\{c_n\}$ of complex numbers, let $\phi_0=1$, $\psi_1=1$, and

$$\phi_\nu=\begin{vmatrix} c_1 & c_2 & \cdots & c_\nu \\ c_2 & c_3 & \cdots & c_{\nu+1} \\ & & \cdots & \\ c_\nu & c_{\nu+1} & \cdots & c_{2\nu-1} \end{vmatrix}, \qquad \psi_\nu=\begin{vmatrix} c_2 & c_3 & \cdots & c_\nu \\ c_3 & c_4 & \cdots & c_{\nu+1} \\ & & \cdots & \\ c_\nu & c_{\nu+1} & \cdots & c_{2\nu-2} \end{vmatrix}$$

$$(\nu=1,2,\cdots) \qquad (\nu=2,3,\cdots).$$

Thus ϕ_ν and ψ_ν are determined once we specify $\{c_n\}$. According to [9, Thm. 7.13, p. 242], the associated continued fraction corresponding to the Maclaurin series of a function analytic at $z=0$ is unique. On the basis of this result, the claim $d_n\neq 0$ for all n, and Satz 3.11 of [21, Satz 3.11, p. 120] (applied to $1+\sqrt{z}\sin\sqrt{z}$) we assert that the d_n in (4.102) are given by

$$d_{2n-1}=-\frac{\psi_{n+1}\phi_{n-1}}{\phi_n\psi_n}, \qquad d_{2n}=-\frac{\phi_{n+1}\psi_n}{\psi_{n+1}\phi_n}, \qquad n\geq 1, \tag{4.105}$$

where $\{c_n\}$ is defined by

$$c_n\equiv\frac{(-1)^{n-1}}{(2n-1)!}, \qquad n\geq 1. \tag{4.106}$$

After extending the continued fraction (4.104), under the assumption that its coefficients are given by (4.105), we obtain the regular δ-fraction

$$\mathsf{K}_z=\frac{z}{1-z}+\frac{z}{1}-\frac{b_1z}{1}+\frac{b_1z}{1}-\frac{b_2z}{1}+\frac{b_2z}{1}$$
$$-\frac{b_3z}{1}+\frac{b_3z}{1}-\frac{b_4z}{1}+\frac{b_4z}{1}-\cdots,$$

where

$$b_1=\frac{1}{6}, \quad b_2=-\frac{7}{10}, \quad b_3=-\frac{11}{686}, \quad b_4=\frac{3857}{2178}$$

and, in general,

$$b_{2n-1}=-\frac{\psi_n\psi_{n+1}}{(\phi_n)^2}, \qquad b_{2n}=\frac{\phi_n\phi_{n+1}}{(\psi_{n+1})^2}, \qquad n\geq 1$$

with $\{c_n\}$ defined by (4.106). Then

$$\mathsf{K}_z\sim\sum_{n=0}^{\infty}\frac{(-1)^n z^{2n+1}}{(2n+1)!}=\sin z.$$

We have not attempted to determine the convergence behavior of K_z.

REFERENCES

[1] D. DIJKSTRA, *A continued fraction expansion for a generalization of Dawson's integral*, Math. Comp., 31 (1977), pp. 503–510.

[2] D. M. DREW AND J. A. MURPHY, *Branch points, M-fractions, and rational approximants generated by linear equations*, J. Inst. Maths Applics., 19 (1977), pp. 169–185.

[3] V. K. DZJADYK, *On the asymptotics of diagonal Padé approximants of the functions* $\sin z$, $\cos z$, $\sinh z$, *and* $\cosh z$, Math. USSR Sbornik, 36 (1980), pp. 231–249.
[4] J. S. FRAME, *The Hankel power sum matrix inverse and the Bernoulli continued fraction*, Math. Comp., 33 (1979), pp. 815–816.
[5] A. O. GEL'FOND, *Calculus of Finite Differences* (authorized English translation of the third Russian edition) Hindustan Publishing Corporation, India, 1971.
[6] J. GILL, *Enhancing the convergence region of a sequence of bilinear transformations*, Math. Scand., 43 (1978), pp. 74–80.
[7] _____, *The use of attractive fixed points in accelerating the convergence of limit-periodic continued fractions*, Proc. Amer. Math. Soc., 47 (1975), pp. 119–126.
[8] T. L. HAYDEN, *Continued fraction approximation to functions*, Numer. Math., 7 (1965), pp. 292–309.
[9] W. B. JONES AND W. J. THRON, *Continued Fractions: Analytic Theory and Applications*, Encyclopedia of Mathematics and Its Applications, Vol. 11, Addison-Wesley, Reading, MA, 1980.
[10] _____, *Sequences of meromorphic functions corresponding to a formal Laurent series*, this Journal, 10 (1979), pp. 1–17.
[11] S. S. KHLOPONIN, *Approximation of functions by continued fractions*, Izvestija VUZ. Matematika, 23 (1979), pp. 37–41. (In Russian.)
[12] A. N. KHOVANSKII, *The Applications of Continued Fractions and Their Generalizations to Problems in Approximation Theory*, Peter Wynn, trans., Noordhoff, Groningen, 1963.
[13] W. LEIGHTON AND W. T. SCOTT, *A general continued fraction expansion*, Bull. Amer. Math. Soc., (1939), pp. 596–605.
[14] A. MAGNUS, *Certain continued fractions associated with the Padé table*, Math. Z., 78 (1962), pp. 361–374.
[15] _____, *Expansion of power series into P-fractions*, Math. Z., 80 (1962), pp. 209–216.
[16] J. H. MCCABE, *A continued fraction expansion, with a truncation error estimate, for Dawson's integral*, Math. Comp., 28 (1974), pp. 811–816.
[17] _____, *A further correspondence property of M-fractions*, Ibid., Comp., 32 (1978), pp. 1303–1305.
[18] J. H. MCCABE AND J. A. MURPHY, *Continued fractions which correspond to power series expansions at two points*, J. Inst. Maths. Applics., 17 (1976), pp. 233–247.
[19] L. M. MILNE-THOMSON, *The Calculus of Finite Differences*, Chapter XVII, Macmillan, London, 1933.
[20] J. A. MURPHY, *Certain rational function approximations to* $(1+x^2)^{-1/2}$, J. Inst. Maths Applics., 7 (1971), pp. 138–150.
[21] O. PERRON, *Die Lehre Von Den Kettenbrüchen*, Band II, Teubner, Stuttgart, 1957.
[22] _____, *Über die Poincarésche lineare Differenzengleichung*, J. Reine Angew. Math., 137 (1909), pp. 6–64.
[23] _____, *Über einen Satz des Herrn Poincaré*, J. Reine Angew. Math., 136 (1909), pp. 17–37.
[24] _____, *Über Summengleichungen und Poincarésche Differenzengleichungen*, Math. Ann., 84 (1921), pp. 1–15.
[25] H. POINCARÉ, *Sur les équations linéaires aux différentielles ordinaires et aux différences finies*, Amer. Math., VII (1885), pp. 203–258.
[26] W. J. THRON, *Some properties of continued fractions* $1+d_0z+K(z/(1+d_n(z)))$, Bull. Amer. Math. Soc., 54 (1948), pp. 206–218.
[27] W. J. THRON AND H. WAADELAND, *Accelerating convergence of limit periodic continued fractions* $K(a_n/1)$, Numer. Math., 34 (1980), pp. 155–170.
[28] _____, *Analytic continuation of functions defined by means of continued fractions*, Math. Scand., 47 (1980), pp. 72–90.
[29] _____, *Convergence questions for limit periodic continued fractions*, to appear.
[30] H. B. VAN VLECK, *On the convergence of algebraic continued fractions, whose coefficients have limiting values*, Trans. Amer. Math. Soc., 5 (1904), pp. 253–262.
[31] H. WAADELAND, *General T-fractions corresponding to functions satisfying certain boundedness conditions*, J. Approx. Theory, 26 (1979), pp. 317–328.
[32] _____, *Limit periodic general T-fractions and holomorphic functions*, J. Approx. Theory, 27 (1979), pp. 329–345.
[33] H. S. HALL, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0010 $01.25/0

A NOTE ON A MULTILINEAR GENERATING FUNCTION FOR THE KONHAUSER BIORTHOGONAL POLYNOMIALS*

H. M. SRIVASTAVA†

Abstract. It is observed, in the present note, that a multilinear generating function for the Konhauser biorthogonal polynomials $Y_n^{\alpha}(x;k)$ and $Z_n^{\alpha}(x;k)$, which is the main result of a recent paper by K. R. Patil and N. K. Thakare [SIAM J. Math. Anal., 9 (1978), pp. 921–923], does not hold true as asserted in the literature. The corrected (and slightly improved) version of this result is given and its relationship with an obvious special case of a known multilinear generating function (due to H. M. Srivastava and J. P. Singhal [Acad. Roy. Belg. Bull. Cl. Sci. (5), 58 (1972), pp. 1238–1247]) for a certain class of generalized hypergeometric polynomials is indicated.

For the Konhauser biorthogonal polynomials $Y_n^{\alpha}(x;k)$ and $Z_n^{\alpha}(x;k)$, Patil and Thakare [1] gave the following multilinear generating function:

$$\sum_{n_1,\cdots,n_r=0}^{\infty}(m+n_1+\cdots+n_r)!\,Y_{m+n_1+\cdots+n_r}^{\alpha+\lambda}(x;k)\prod_{i=1}^{r}\left\{\frac{Z_{n_i}^{\beta_i}(y_i;s)u_i^{n_i}}{(1+\beta_i)_{sn_i}}\right\}$$

$$=e^{x}\Delta_r^{(-(\alpha+1)/k-m)(1+\lambda)}\sum_{l=0}^{\infty}\frac{1}{l!}\left(\frac{\alpha+\lambda+l+1}{k}\right)_m\left(-\frac{x}{\Delta_r^{(\lambda+1)/k}}\right)^{l}$$

$$\cdot\Psi_2^{r}\left[\frac{\alpha+\lambda+l+1}{k}+m;\,(1+\beta_1)^s,\cdots,(1+\beta_r)^s;\right.$$

$$\left.-\frac{y_1^s u_1}{s\Delta_r^{\lambda+1}},\cdots,-\frac{y_r^s u_r}{s\Delta_r^{\lambda+1}}\right], \tag{1}$$

$$\Delta_r\equiv 1-\sum_{i=1}^{r}u_i,$$

where Ψ_2^n denotes a confluent hypergeometric function of n variables, and $(\lambda)_n=\Gamma(\lambda+n)/\Gamma(\lambda)$. Their derivation of (1) was based rather heavily upon the differential operator

$$\theta_{x,\lambda}\equiv x^k(\lambda+xD_x),\qquad D_x=\frac{d}{dx}, \tag{2}$$

where λ is a constant. In our attempt to give a direct proof of (1), *without* using the differential operator $\theta_{x,\lambda}$, we were led to the interesting fact that the multilinear generating function (1) does not hold true as asserted by Patil and Thakare [1]. Indeed, in terms of a hypergeometric function of $r+1$ variables (see Srivastava and Daoust [2,

* Received by the editors April 4, 1981, and in revised form November 2, 1981. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A-7353.

† Department of Mathematics, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada.

p. 454]), we are thus led to the multilinear generating function

(3)

$$\sum_{n_1,\cdots,n_r=0}^{\infty}(m+n_1+\cdots+n_r)!\,Y^{\alpha}_{m+n_1+\cdots+n_r}(x;k)\prod_{i=1}^{r}\left\{\frac{Z^{\beta_i}_{n_i}(y_i;s_i)u_i^{n_i}}{(1+\beta_i)_{s_in_i}}\right\}$$

$$=\left(\frac{\alpha+1}{k}\right)_m e^{x}\Delta_r^{-m-(\alpha+1)/k}$$

$$F^{1:0;\cdots;0}_{0:1;\cdots;1}\left(\begin{matrix}[m+(\alpha+1)/k:1/k,1,\cdots,1]:\text{———————}\\ \text{———————}:[(\alpha+1)/k:1/k]\end{matrix};\right.$$

$$\left.\begin{matrix}\text{———}\\ [1+\beta_1:s_1]\end{matrix};\cdots;\begin{matrix}\text{———}\\ [1+\beta_r:s_r]\end{matrix};-\frac{x}{\Delta_r^{1/k}},-\frac{y_1^{s_1}u_1}{\Delta_r},\cdots,-\frac{y_r^{s_r}u_r}{\Delta_r}\right),$$

whose special case, when $s_i=s$, $i=1,\cdots,r$, differs markedly from the Patil–Thakare result (1) with, of course, α replaced trivially by $\alpha-\lambda$. Formula (3) is the *corrected* (and slightly improved) version of the erroneous result (1).

Since $s_1,\cdots,s_r$ are (by definition) positive integers, the multilinear generating function (3), with $k=1$, can be derived easily as a special case of a known result [3, eq. (24), p. 1244] with x replaced on both sides by x/ν and $\nu\to\infty$. We omit the details involved.

It may be of interest to remark that the only situation in which the right-hand side of (3) can be expressed in terms of the confluent hypergeometric Ψ_2^n function occurring in (1) is when $k=s_1=\cdots=s_r=1$, in which case (3) at once yields, for the Laguerre polynomials $L_n^{(\alpha)}(x)=Y_n^{\alpha}(x;1)=Z_n^{\alpha}(x;1)$,

$$\sum_{n_1,\cdots,n_r=0}^{\infty}(m+n_1+\cdots+n_r)!\,L^{(\alpha)}_{m+n_1+\cdots+n_r}(x)\prod_{i=1}^{r}\left\{\frac{L^{(\beta_i)}_{n_i}(y_i)u_i^{n_i}}{(1+\beta_i)_{n_i}}\right\}$$

$$=(\alpha+1)_m e^{x}\Delta_r^{-\alpha-m-1}\Psi_2^{r+1}\left[\alpha+m+1;\alpha+1,\beta_1+1,\cdots,\beta_r+1;-\frac{x}{\Delta_r},-\frac{y_1u_1}{\Delta_r},\cdots,-\frac{y_ru_r}{\Delta_r}\right] \tag{4}$$

or, equivalently,

$$\sum_{n_1,\cdots,n_r=0}^{\infty}(m+n_1+\cdots+n_r)!\,L^{(\alpha)}_{m+n_1+\cdots+n_r}(x)\prod_{i=1}^{r}\left\{\frac{L^{(\beta_i)}_{n_i}(y_i)u_i^{n_i}}{(1+\beta_i)_{n_i}}\right\}$$

$$=e^{x}\Delta_r^{-\alpha-m-1}\sum_{l=0}^{\infty}\frac{(\alpha+l+1)_m}{l!}\left(-\frac{x}{\Delta_r}\right)^{l}$$

$$\cdot\Psi_2^{r}\left[\alpha+l+m+1;\beta_1+1,\cdots,\beta_r+1;-\frac{y_1u_1}{\Delta_r},\cdots,-\frac{y_ru_r}{\Delta_r}\right], \tag{5}$$

where Δ_r is defined with (1).

The multilinear generating function (4) is due to Srivastava and Singhal [3, eq. (5), p. 1239] who proved it in two different ways; it was this result of Srivastava and Singhal [3] which Patil and Thakare [1] had set out to generalize for the Konhauser biorthogonal polynomials $Y_n^{\alpha}(x;k)$ and $Z_n^{\alpha}(x;k)$.

Finally, we remark that the errors in the derivation by Patil and Thakare [1, p. 922] are due, for example, to a misinterpretation of a well-known hypergeometric generating function for the Konhauser polynomials $Z_n^{\alpha}(x;k)$ and several improper uses of such operational formulas as

$$\theta_{x,\lambda}^{n}\{x^{\mu}\} = x^{\mu+kn}k^{n}\left(\frac{\lambda+\mu}{k}\right)_n, \qquad n=0,1,2,\cdots, \tag{6}$$

which is an easy consequence of the definition (2).

REFERENCES

[1] K. R. Patil and N. K. Thakare, *Multilinear generating function for the Konhauser biorthogonal polynomial sets*, this Journal, 9 (1978), pp. 921–923.

[2] H. M. Srivastava and M. C. Daoust, *Certain generalized Neumann expansions associated with the Kampé de Fériet function*, Nederl. Akad. Wetensch. Proc. Ser. A, 72 = Indag. Math., 31 (1969), pp. 449–457.

[3] H. M. Srivastava and J. P. Singhal, *Some formulas involving the products of several Jacobi or Laguerre polynomials*, Acad. Roy. Belg. Bull. Cl. Sci. (5), 58 (1972), pp. 1238–1247.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0011 $01.25/0

RECIPROCAL POWER SUMS OF DIFFERENCES OF ZEROS OF SPECIAL FUNCTIONS*

S. AHMED† AND MARTIN E. MULDOON‡

Abstract. We derive formulas for sums of the form

$$\sum_{\substack{k=1\\k\neq j}}^{\infty}(z_j-z_k)^{-m} \qquad (m,j=1,2,3,\cdots),$$

where $\{z_k\}$ is the (finite or infinite) sequence of (complex) zeros of an appropriate solution of a second order linear differential equation. In special cases our results reduce to some of those obtained by T. J. Stieltjes, F. Calogero, S. Ahmed, M. L. Mehta, K. M. Case and others.

1. Introduction. Recently there has been a renewed interest in results concerning sums of the form

$$S_{m,j}=\sum_{k}{}'(z_j-z_k)^{-m}, \tag{1.1}$$

$m=1,2,\cdots$, $j=1,2,\cdots$, where $\{z_k\}$ is the finite or infinite sequence of (complex) zeros of an appropriate solution of a second order linear differential equation. *Here and in what follows the prime on a summation sign indicates that the singular term ($k=j$ in (1.1)) must be omitted.* In the case $m=1$ such results were discovered for zeros of various polynomials by Stieltjes [20] (see also [22, pp. 140–142]) and were related by him to the interpretation of these zeros as equilibrium positions for certain one-dimensional electrostatic problems. A typical and very simple example is the set of equations

$$\sum_{k=1}^{n}{}'(x_j-x_k)^{-1}=x_j, \qquad j=1,\cdots,n$$

for the zeros of the Hermite polynomial $H_n(x)$.

F. Calogero [9] initiated a study of the corresponding relations for zeros of the Bessel function $J_p(x)$ in connection with a study of integrable many-body problems, and corresponding results for a host of other special functions have been given in a series of papers by Calogero, S. Ahmed and M. Bruschi. Ahmed et al. [4] have summarized a good deal of this work and have connected it with the theory of certain matrices having integral eigenvalues. Some of the results have application to finding bounds for zeros of special functions [5], [6], [1] while others are used in the study of the asymptotic density of zeros of orthogonal polynomials [11], [12] and in the study of the monotonicity of the zeros with respect to certain parameters [21], [22, pp. 123–124].

*Received by the editors May 27, 1981, and in revised form December 4, 1981. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

†Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3. Present address, Department of Mathematics, Kings College, Wilkes-Barre, Pennsylvania 18711. Some of this work was done during periods when this author was at the Department of Physics, University of Alberta and the Department of Mathematics, University of Toronto.

‡Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3. Some of this work was done while this author was visiting the Department of Mathematics, University of Dundee.

The sums (1.1) are closely related to formulas (the so called "sum rules") for sums of the form

$$\sum_k z_k^{-m};$$

see [5], [9], [10], [13], [14].

There have been two attempts to unify results on sums such as (1.1) for various special functions. M. L. Mehta [19] has given a procedure for finding the sums $S_{m,j}$ $(m=1,2,\cdots)$ successively and K. M. Case [13], [14] has discussed these and other sums in more detail for solutions of

$$g_2(x)y''+g_1(x)y'+g_0(x)y=0. \tag{1.2}$$

All of Mehta's and most of Case's results are concerned with polynomial solutions, and the coefficients in (1.2) are supposed to be polynomials of low degree. In the case where the solutions are entire functions rather than polynomials the method introduced in [9] and described in [4] and the method used by Case [13], [14] involve delicate questions concerning the validity of certain interchanges in the order of summation of double series. While there is no doubt as to the correctness of the final results, it is by no means clear to us how all of these interchanges are to be justified.

The purpose of this paper is to re-examine the derivation of sums of the form (1.1) and

$$\sum_j (z_j-\alpha)^{-1}$$

where the z_j are zeros of an appropriate solution of a differential equation

$$y''+P(z)y'+Q(z)y=0, \tag{1.3}$$

having a singularity at α. In the case of several singular points it is only in exceptional circumstances that an entire solution of (2.1) exists. Nevertheless, our general results can be used to obtain most of the earlier results known to us on the sums (1.1). Our method is based on the introduction of the function $y_j(z)$ and the exploitation of its relation to $y(z)$ given in (2.7) and (5.12). This shortens many of the earlier arguments and avoids the problems, referred to above, concerning the interchange of orders of summation.

2. The sums $S_{1,j}$. Here we suppose that y is an entire function which satisfies

$$y''+P(z)y'+Q(z)y=0 \tag{2.1}$$

where P and Q are meromorphic, and that y has simple zeros at the nonzero points $z_1, z_2, \cdots$. We suppose that none of the z's coincides with a singularity of P or Q and that

$$\sum |z_k|^{-1}<\infty. \tag{2.2}$$

We remark that, in some cases, (2.2) will be guaranteed by the forms of P and Q; see, for example [16, Thm. 4.6.3, pp. 137–138] together with [23, 8.22, p. 249]. We suppose, further, that y has a Weierstrass product representation given by

$$y(z)=e^{g(z)}\prod_{k=1}^{\infty}\left(1-\frac{z}{z_k}\right), \tag{2.3}$$

where

$$g(z)=n\log z+h(z),$$

n being a nonnegative integer and h an entire function. We introduce the functions

$$y_j(z)=e^{g(z)}\prod_{\substack{k=1\\k\neq j}}^{\infty}\left(1-\frac{z}{z_k}\right),\qquad j=1,2,\cdots. \tag{2.4}$$

Thus

$$y(z)=\left(1-\frac{z}{z_j}\right)y_j(z).$$

Hence,

$$y'(z)=-z_j^{-1}y_j(z)+\left(1-\frac{z}{z_j}\right)y_j'(z) \tag{2.5}$$

and

$$y''(z)=-2z_j^{-1}y_j'(z)+\left(1-\frac{z}{z_j}\right)y_j''(z). \tag{2.6}$$

This shows that

$$\frac{y''(z_j)}{y'(z_j)}=2\frac{y_j'(z_j)}{y_j(z_j)}. \tag{2.7}$$

On the other hand, if we differentiate (2.4) logarithmically (this is justified, e.g., by [18, pp. 14–15]) we get

$$\frac{y_j'(z)}{y_j(z)}=\sum_{\substack{k=1\\k\neq j}}^{\infty}(z-z_k)^{-1}+g'(z), \tag{2.8}$$

whenever the right-hand side has meaning, and

$$\frac{y_j'(z_j)}{y_j(z_j)}=S_{1,j}+g'(z_j).$$

Comparing this with (2.7) we get

$$S_{1,j}=\frac{1}{2}\frac{y''(z_j)}{y'(z_j)}-g'(z_j). \tag{2.9}$$

From (2.1) we have

$$y''(z_j)+P(z_j)y'(z_j)=0,$$

so (2.9) gives finally

$$S_{1,j}=-\frac{1}{2}P(z_j)-g'(z_j). \tag{2.10}$$

In the important special case where $g(z)$ is constant, this becomes

$$S_{1,j}=-\frac{1}{2}P(z_j).$$

3. The sum $\Sigma(z_k-\alpha)^{-1}$. We suppose now that the hypotheses of §2 hold and that α is a singular point of the differential equation (2.1) such that the functions $g'(z)$, $g''(z)$ and $Q(z)/P(z)$ have finite limits as $z\to\alpha$. Logarithmic differentiation of (2.3), justified by [18, pp. 14–15], leads to

$$\frac{y''(z)}{y(z)}=\left[\sum_k (z-z_k)^{-1}+g'(z)\right]^2-\sum_k (z-z_k)^{-2}+g''(z).$$

In the neighbourhood of $z=\alpha$, the differential equation (2.1) may be written

$$(z-\alpha)\frac{y''}{y}+(z-\alpha)P(z)\frac{y'}{y}+(z-\alpha)Q(z)=0.$$

Letting $z\to\alpha$ we get

$$\lim_{z\to\alpha}(z-\alpha)P(z)\left\{\sum_k (\alpha-z_k)^{-1}+g'(\alpha)\right\}+\lim_{z\to\alpha}(z-\alpha)Q(z)=0$$

or

$$\sum_k (z_k-\alpha)^{-1}=\lim_{z\to\alpha}\frac{Q(z)}{P(z)}+g'(\alpha). \tag{3.1}$$

In case $g(z)$ is constant this becomes

$$\sum_k (z_k-\alpha)^{-1}=\lim_{z\to\alpha}\frac{Q(z)}{P(z)}.$$

4. Some special cases. We suppose now that equation (2.1) has n singular points $\alpha_1,\cdots,\alpha_n$ and that

$$P(z)=p_0(z)+\sum_{i=1}^{n} p_i(z-\alpha_i)^{-1},$$

$$Q(z)=q_0(z)+\sum_{i=1}^{n} q_i(z-\alpha_i)^{-1},$$

where p_0 and q_0 are entire functions and $p_i\neq 0$, $i=1,\cdots,n$. It then follows from the differential equation (2.1) that no *simple* zero of y can occur at one of the α_i. The other assumptions of §2 are supposed to hold here. We see that the results (2.10) and (3.1) become

$$\sum_{k=1}^{\infty}{}' (z_j-z_k)^{-1}=-\frac{1}{2}p_0(z_j)-\frac{1}{2}\sum_{i=1}^{n} p_i(z_j-\alpha_i)^{-1}-g'(z_j), \qquad j=1,2,\cdots, \tag{4.1}$$

and

$$\sum_{k=1}^{\infty} (z_k-\alpha_i)^{-1}=\frac{q_i}{p_i}+g'(\alpha_i), \qquad i=1,\cdots,n. \tag{4.2}$$

An example in the case of one singular point is provided by the function

$$y(z)=\Gamma(p+1)2^p z^{-p/2}J_p(z^{1/2})=\prod_{k=1}^{\infty}\left(1-\frac{z}{z_k}\right)$$

which satisfies

$$y''+(1+p)z^{-1}y'+\frac{1}{4}z^{-1}y=0,$$

so that (4.1) and (4.2) give the result of Calogero [9]

$$\sum_{k=1}^{\infty}{}'(z_j-z_k)^{-1}=-\frac{1}{2}(1+p)z_j^{-1}, \qquad j=1,2,\cdots,$$

and the well known

$$\sum_{k=1}^{\infty} z_k^{-1}=[4(1+p)]^{-1}.$$

Ahmed and Calogero [5] have considered the function

$$y(z)=(p+q)^{-1}\Gamma(p+1)\left(\frac{z}{4}\right)^{-p/2}\left[qJ_p(z^{1/2})+z^{1/2}J_p'(z^{1/2})\right]$$

and noted that

$$\varphi(z)=y[(p^2-q^2)z]$$

satisfies (2.1) with

$$P(z)=(p+1)z^{-1}-(z-1)^{-1},$$
$$Q(z)=\frac{1}{4}(p-q)(p+q+2)z^{-1}+\frac{1}{2}(q-p)(z-1)^{-1}.$$

We suppose that $p>-1$, so that by Dixon's theorem [24, p. 480], φ has only simple zeros (other than $x=0$) and so doesn't vanish at 1. Now

$$y(z)=2(p+q)^{-1}\Gamma(p+1)2^p z^{1-(p+q)/2}\frac{d}{dz}\left\{z^{q/2}J_p(z^{1/2})\right\},$$

so y is an entire function of order $\frac{1}{2}$ and the Hadamard factorization theorem [23, p. 250] shows that if the zeros of $\varphi(z)$ are $z_1, z_2, \cdots$, we have

$$\varphi(z)=\prod_{k=1}^{\infty}\left(1-\frac{z}{z_k}\right).$$

Thus the results

$$\sum_{k=1}^{\infty}{}'(z_k-z_j)^{-1}=\frac{1}{2}(p+1)z_j^{-1}+\frac{1}{2}(1-z_j)^{-1}, \qquad j=1,2,\cdots,$$
$$\sum_{j=1}^{\infty} z_j^{-1}=\frac{1}{4}\frac{(p+q+2)(p-q)}{p+1},$$

and

$$\sum_{j=1}^{\infty}(z_j-1)^{-1}=\frac{1}{2}(p-q)$$

of [5] follow from (4.1) and (4.2). We have not considered the case $\nu\leq -1$, but it seems likely that the results can still be gotten from our general results in this case.

The results of this section are applicable also to the Lamé equation in algebraic form [15, p. 56]:

$$\frac{d^2\Lambda}{dx^2}+\frac{1}{2}\left(\frac{1}{x}+\frac{1}{x-1}+\frac{1}{x-k^{-2}}\right)\frac{d\Lambda}{dx}+\frac{hk^{-2}-(n)(n+1)x}{4x(x-1)(x-k^{-2})}\Lambda=0. \tag{4.3}$$

With the notation of [8, Chapt. 9] we let $z_1,\cdots,z_m$ be the zeros of the Lamé polynomial $E_{2N}^m(z)$ of the first species. We get, from (4.1) and (4.2),

$$\sum_{l=1}^{m}(z_j-z_l)^{-1}=-\frac{1}{4}\left\{z_j^{-1}+(z_j-1)^{-1}+\left(z_j-k^{-2}\right)^{-1}\right\},\qquad j=1,\cdots,m, \tag{4.4}$$

$$\sum_{j=1}^{m}z_j^{-1}=\frac{h}{2}, \tag{4.5}$$

$$\sum_{j=1}^{m}(z_j-1)^{-1}=\frac{n(n+1)k^2-h}{2(1-k^2)}, \tag{4.6}$$

and

$$\sum_{j=1}^{m}\left(z_j-k^{-2}\right)^{-1}=\frac{hk^2-n(n+1)k^2}{2(1-k^2)}. \tag{4.7}$$

The result (4.4), in a different notation, is given by Arscott [8, Exer. 13, p. 233], who bases it on the work of Stieltjes [20]. The sum rule (4.5) was noted by Case [14]. Of course (4.5), (4.6), and (4.7) may also be found by examining the series expansions of the appropriate solution of (4.3) about the points 0, 1 and k^{-2} respectively.

Our general results may, of course, be applied to other solutions of the Lamé equation. They may also be applied to the generalized Lamé equation [17, p. 496].

5. Higher order sums. Here we consider the sums

$$S_{m,j}=\sum_{k=1}^{\infty}{}'(z_j-z_k)^{-m}\qquad(j,m=1,2,\cdots). \tag{5.1}$$

with the notation and assumptions of §2. Our principal result expresses $S_{m,j}$ in terms of a determinant involving the coefficients in the equation (2.1) and their derivatives.

THEOREM 5.1.

$$S_{m,j}=[(m-1)!]^{-1}\left\{\det\left[a_{r,s}(z_j)\right]_{r,s=1,\cdots,m}+(-1)^m g^{(m)}(z_j)\right\},$$

where

$$a_{r,s}(z_j)=\begin{cases}(s+1)^{-1}A_{s+1}(z_j), & r=1,\\ (s-r+2)^{-1}\binom{s-1}{r-2}A_{s-r+2}(z_j), & 1<r<s+1,\\ 1, & r=s+1,\\ 0, & r>s+1.\end{cases}$$

and $A_2=-P$, $B_2=-Q$,

$$A_{n+1}=A_2A_n+A_n'+B_n,\qquad B_{n+1}=B_2A_n+B_n',\qquad n=2,3,\cdots.$$

COROLLARY. *For the first few values of m, this implies*

$$S_{1,j}=\sum_{k=1}^{\infty}{}'(z_j-z_k)^{-1}=-\frac{1}{2}P(z_j)-g'(z_j),$$

$$3S_{2,j}=3\sum_{k=1}^{\infty}{}'(z_j-z_k)^{-2}=-\frac{1}{4}\left[P(z_j)\right]^2+P'(z_j)+Q(z_j)+3g''(z_j),$$

$$8S_{3,j}=8\sum_{k=1}^{\infty}{}'(z_j-z_k)^{-3}=P(z_j)P'(z_j)-P''(z_j)-2Q'(z_j)-4g'''(z_j),$$

etc.

Remark. Mehta [19] has a result (for polynomial solutions only) which gives $S_{m,j}$ implicitly in terms of determinants involving the coefficients in the equation and their derivatives.

Proof of Theorem 5.1. We have

$$S_{m,j}=S_{m,j}(z_j), \tag{5.2}$$

where

$$S_{m,j}(z)=\sum_{\substack{k=1\\k\neq j}}^{\infty}(z-z_k)^{-m}, \tag{5.3}$$

and this gives

$$S'_{m,j}(z)=-mS_{m+1,j}(z), \qquad m=1,2,\cdots. \tag{5.4}$$

Now we introduce the notation

$$w_n(z)=\frac{y_j^{(n)}(z)}{y_j(z)}.$$

Clearly, we have

$$w_n'=w_{n+1}-w_1w_n. \tag{5.5}$$

We define a sequence of determinants $\Delta_{m,j}(z)$ $(m=1,2,\cdots)$ by

$$\Delta_{m,j}(z)=\det\left[\alpha_{r,s}(z)\right]_{r,s=1,\cdots,m},$$

where

$$\alpha_{r,s}(z)=\begin{cases} w_s(z), & r=1,\\ \binom{s-1}{r-2}w_{s-r+1}(z), & 1<r<s+1,\\ 1, & r=s+1,\\ 0, & r>s+1.\end{cases}$$

Expanding the determinant by its last column we get

$$(-1)^{m+1}\Delta_{m,j}(z)=w_m(z)+\sum_{k=0}^{m-2}(-1)^{k+1}\binom{m-1}{k}w_{m-k-1}(z)\Delta_{k+1,j}(z). \tag{5.6}$$

Next we show that

$$\Delta'_{m,j}(z)=-\Delta_{m+1,j}(z), \qquad m=1,2,\cdots. \tag{5.7}$$

Equation (5.7) clearly holds for $m=1$. We suppose that it holds for $m=1,\cdots,N-1$. We wish to show

$$\Delta'_{N,j}(z)=-\Delta_{N+1,j}(z). \tag{5.8}$$

We have, from (5.6),

$$\begin{aligned}(-1)^{N+1}\Delta'_{N,j}=w'_N+\sum_{k=0}^{N-2}(-1)^{k+1}\binom{N-1}{k}w'_{N-k-1}\Delta_{k+1,j}\\ +\sum_{k=0}^{N-2}(-1)^{k+1}\binom{N-1}{k}w_{N-k-1}\Delta'_{k+1,j}.\end{aligned} \tag{5.9}$$

We use (5.5) in the first term and the first sum on the right-hand side here. In the second sum we use our inductive hypothesis that (5.7) holds for $m=1,\cdots,N-1$. Some standard manipulations with series and binomial coefficients then show that the right-hand side is $(-1)^N\Delta_{N+1,j}(z)$. Thus (5.8) holds and (5.7) has been proved.

Next we have

$$(n-1)!S_{n,j}(z)=\Delta_{n,j}(z)+(-1)^n g^{(n)}(z), \qquad n=1,2,\cdots. \tag{5.10}$$

This is clear for $n=1$, and for the other values it follows from the recurrence relations (5.4) and (5.7) satisfied by $S_{m,j}(z)$ and $\Delta_{m,j}(z)$ respectively.

In order to prove the theorem it remains to show that $\alpha_{r,s}(z_j)=a_{r,s}(z_j)$, i.e., that

$$w_s(z_j)=(s+1)^{-1}A_{s+1}(z_j), \qquad s=1,2,\cdots. \tag{5.11}$$

In order to see this, we recall (2.5) and (2.6) and remark that successive differentiation gives

$$y^{(n)}(z)=-nz_j^{-1}y_j^{(n-1)}(z)+\left(1-\frac{z}{z_j}\right)y_j^{(n)}(z), \qquad n=1,2,\cdots,$$

and so

$$w_s(z_j)=(s+1)^{-1}\frac{y^{(s+1)}(z_j)}{y'(z_j)}. \tag{5.12}$$

On the other hand, successive differentiation of (2.1) leads to

$$y^{(s+1)}(z_j)=A_{s+1}(z_j)y'(z_j). \tag{5.13}$$

Comparing (5.12) and (5.13) we get (5.11), so the proof of the theorem is complete. □

Theorem 5.1 may be applied to all of the examples considered in §4. In this way, for example, we get the results of Calogero [10] on sums of the form (5.1) for zeros of $z^{p/2}J_p(z^{1/2})$.

6. Confluent hypergeometric and related functions. In the foregoing it was assumed that

$$\sum|z_j|^{-1}<\infty, \tag{6.1}$$

and this was necessary in order to be able to deal with

$$\prod_{k=1}^{\infty}\left(1-\frac{z}{z_k}\right)$$

and related products. This precludes results on functions whose zeros grow too slowly in modulus for (6.1) to hold but for which

$$\sum |z_k|^{-2} < \infty. \tag{6.2}$$

Of course we cannot expect to get results of the kind considered in §2 and §3 for such functions, but one may still expect to get results of the kind considered in §5. Indeed Ahmed [2], [3] has discussed these questions for the zeros of the confluent hypergeometric function ${}_1F_1(a, c; z)$ by starting with sums of the form

$$\sum_{j=1}^{\infty}{}' x_j^{-1}(x_j - x_k)^{-1},$$

which are clearly convergent under the assumption (6.2), and going on to find the sums (5.1) for $m = 2, 3, \cdots$. Here the results of [2], [3] are put in a more general setting.

We consider an entire function y which satisfies (2.1) and whose zeros are such that (6.2) holds. Otherwise the assumptions of §2 are supposed to hold. We suppose that the resulting Weierstrass representation of y has the form

$$y(z) = e^{g(z)} \prod_{k=1}^{\infty} \left(1 - \frac{z}{z_k}\right) e^{z/z_k} \tag{6.3}$$

where $g(z)$ is as in §2. In analogy with §§2 and 5 we let y_j be the corresponding product with the factor $(1 - z/z_j)$ omitted, i.e.,

$$y_j(z) = e^{g(z)} e^{z/z_j} \prod_{\substack{k=1 \\ k \neq j}}^{\infty} \left(1 - \frac{z}{z_k}\right) e^{z/z_k}. \tag{6.4}$$

We have

$$\frac{y_j'(z)}{y_j(z)} = z \sum_{\substack{k=1 \\ k \neq j}}^{\infty} z_k^{-1}(z - z_k)^{-1} + z_j^{-1} + g'(z). \tag{6.5}$$

The formula for y_j''/y_j is

$$\frac{y_j''(z)}{y_j(z)} - \left[\frac{y_j'(z)}{y_j(z)}\right]^2 = -\sum_{\substack{k=1 \\ k \neq j}}^{\infty} (z - z_k)^{-2} + g''(z), \tag{6.6}$$

which is just (5.10) for $n = 2$. The relation between the derivatives of y and y_j is (5.12), the same as before.

It is then easy to see that Theorem 5.1 holds *provided we interpret* $S_{1,j}$ *as*

$$z_j^{-1} + z_j \sum_{k=1}^{\infty}{}' z_k^{-1}(z_j - z_k)^{-1},$$

the other $S_{m,j}$ having the same meaning as in §5.

We now take

$$y(z) = e^{-az/c}\, {}_1F_1(a, c; z)$$

where a and c are chosen so that the zeros of this function satisfy (6.2). For this it is sufficient that a be complex and c real, but c not be zero or a negative integer; see

[7, footnote 3]. We can show y is an entire function satisfying (2.1) with

$$P(z)=\frac{(2a-c)}{c}+cz^{-1},$$

$$Q(z)=\frac{a(a-c)}{c^2}$$

and

$$y(z)=\prod_{k=1}^{\infty}\left(1-\frac{z}{z_k}\right)e^{z/z_k},$$

where $\{z_k\}$ are the zeros of ${}_1F_1(a,c;x)$. Thus

$$x_j^{-1}+x_j\sum_{k=1}^{\infty}{}' x_k^{-1}(x_j-x_k)^{-1}=\frac{c-2a}{(2c)}-c(2x_j)^{-1}, \qquad j=1,2,\cdots. \tag{6.7}$$

This agrees with the result obtained in [2]. Moreover, for the first few values of m, Theorem 5.1 gives

$$12x_j^2\sum_{k=1}^{\infty}{}'(x_j-x_k)^{-2}=-x_j^2+2(c-2a)x_j-c(c+4), \qquad j=1,2,\cdots, \tag{6.8}$$

$$8x_j^3\sum_{k=1}^{\infty}{}'(x_j-x_k)^{-3}=(c-2a)x_j-c(c+2), \tag{6.9}$$

etc. in agreement with [3, (6), (7), etc.].

Acknowledgments. We thank Professor B. D. Sleeman for pointing out that the result (4.4) occurs in reference [8]. We are grateful to a referee for comments leading to a more accurate title and introduction.

REFERENCES

[1] S. AHMED, *Systems of nonlinear equations for the zeros of Hermite polynomials*, Lett. Nuovo Cimento, 22 (1978), pp. 367–370.

[2] _____, *On the zeros of confluent hypergeometric functions.* I: *An infinite system of nonlinear equations*, Lett. Nuovo Cimento, 25 (1979), pp. 520–522.

[3] _____, *On the zeros of confluent hypergeometric functions.* II: *Higher order nonlinear equations and sum rules*, Lett. Nuovo Cimento, 25 (1979), pp. 523–526.

[4] S. AHMED, M. BRUSCHI, F. CALOGERO, M. A. OLSHANETSKY AND A. M. PERELOMOV, *Properties of the zeros of the classical polynomials and of the Bessel functions*, Nuovo Cimento, 49B (1979), pp. 173–199.

[5] S. AHMED AND F. CALOGERO, *On the zeros of Bessel functions.* III, Lett. Nuovo Cimento, 21 (1978), pp. 311–314.

[6] _____, *On the zeros of Bessel functions.* IV, Lett. Nuovo Cimento, 21 (1978), pp. 531–534.

[7] S. AHMED AND M. E. MULDOON, *On the zeros of confluent hypergeometric functions.* III: *Characterizations by means of nonlinear equations*, Lett. Nuovo Cimento, 29 (1980), pp. 353–358.

[8] F. M. ARSCOTT, *Periodic Differential Equations. An Introduction to Mathieu, Lamé and Allied Functions*, Pergamon Press, Oxford, 1964.

[9] F. CALOGERO, *On the zeros of Bessel functions*, Lett. Nuovo Cimento, 20 (1977), pp. 254–256.

[10] _____, *On the zeros of Bessel functions.* II, Lett. Nuovo Cimento, 20 (1977), pp. 476–478.

[11]. _____, *Asymptotic behaviour of the zeros of the Jacobi polynomial* $P_n^{(at,bt)}(x)$ *as* $t\to\infty$ *and limit relations of these polynomials with Hermite polynomials*, Lett. Nuovo Cimento, 23 (1978), pp. 167–168.

[12] F. CALOGERO AND A. M. PERELOMOV, *Asymptotic density of the zeros of Hermite polynomials of diverging order and related properties of certain singular integral operators*, Lett. Nuovo Cimento, 23 (1978), pp. 650–652.
[13] K. M. CASE, *Sum rules for zeros of polynomials.* I, J. Math. Phys., 21 (1980), pp. 702–708.
[14] _____, *Sum rules for zeros of polynomials.* II, J. Math. Phys., 21 (1980), pp. 709–714.
[15] A. ERDÉLYI ET AL., *Higher Transcendental Functions*, vol. 3, McGraw-Hill, New York, 1955.
[16] E. HILLE, *Differential Equations in the Complex Domain*, John Wiley, New York, 1976.
[17] E. L. INCE, *Ordinary Differential Equations*, Longmans, London, 1927.
[18] K. KNOPP, *The Theory of Functions*, part 2, Dover, New York, 1947.
[19] M. L. MEHTA, *Properties of the zeros of a polynomial satisfying a second order linear partial differential equation*, Lett. Nuovo Cimento, 26 (1979), pp. 361–362.
[20] T. J. STIELTJES, *Sur certains polynômes qui vérifient une équation differentielle linéaire du second ordre et sur la theorie des fonctions de Lamé*, Acta Math., 6 (1885), pp. 321–326 (Oeuvres Complêtes I, pp. 434–439).
[21] _____, *Sur les racines de l'équation* $X_n=0$, Acta Math., 9 (1886), pp. 385–400 (Oeuvres Complêtes II, pp. 73–88).
[22] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, 4th ed., American Mathematical Society, Providence, RI, 1975.
[23] E. C. TITCHMARSH, *The Theory of Functions*, 2nd ed., Oxford Univ. Press, London, 1939.
[24] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, London, 1944.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0012 $01.25/0

INEQUALITIES AND APPROXIMATIONS FOR ZEROS OF BESSEL FUNCTIONS OF SMALL ORDER*

ANDREA LAFORGIA† AND MARTIN E. MULDOON‡

Abstract. Let $j_{\nu k}$ and $c_{\nu k}$ denote the kth positive zeros of $J_\nu(x)$ and of the general cylinder function $C_\nu(x)$ respectively. Using a result of Á. Elbert (Studia Sci. Math. Hungar., 12 (1977)), pp. 81–88 on the concavity of $j_{\nu k}$ as a function of ν, we prove various inequalities for this function. We find the first three terms in the power series expansion of $c_{\nu k}$ as a function of ν and give the numerical values of the coefficients in the case of $j_{\nu 1}$.

1. Introduction. For $\nu \geq 0$, we use $j_{\nu k}$ and $c_{\nu k}$ to denote the kth positive zeros of the Bessel function $J_\nu(x)$ and of the general cylinder function

$$C_\nu(x) = \cos\alpha J_\nu(x) - \sin\alpha Y_\nu(x),$$

where α is fixed, $0 \leq \alpha < \pi$. In particular, if $\alpha = 0$, $c_{\nu k} = j_{\nu k}$. The definitions may be extended to negative values of ν in such a way that $c_{\nu k}$ varies continuously with ν, $c_{\nu k} \to 0$ when $\nu \to \alpha/\pi - k$ and on the interval

$$\frac{\alpha}{\pi} - k < \nu < \frac{\alpha}{\pi} - k + 1,$$

$c_{\nu k}$ is the first positive zero of $C_\nu(x)$; see [11, p. 508], [2]. The general behaviour of $j_{\nu k}$ and $c_{\nu k}$ as functions of ν may be seen from the graphs in [11, p. 510].

Our first purpose here is to prove some inequalities for $j_{\nu k}$ which are particularly sharp for ν close to 0 and which extend or complement some of those already known [3], [4], [6], [7]. Our second purpose is to find the first three terms in the power series expansion of $c_{\nu k}$ as a function of ν. In particular, we find

$$j_{\nu 1} = j_{01} + 1.542889743\,\nu - 0.175493592\,\nu^2 + O(\nu^3), \tag{1.1}$$

where the coefficients have been rounded off in the ninth decimal place. Such results, useful for very small ν, can be thought of as complementing those of F. G. Tricomi ([10]; see also [8, Exer. 6.4, p. 408]), e.g.,

$$j_{\nu k} = \nu + a_k \nu^{1/3} + b_k \nu^{-1/3} + O(\nu^{-1}), \qquad \nu \to \infty, \tag{1.2}$$

which are useful for large ν.

2. Lower bounds for $j_{\nu k}$. Our chief tool here is a result of Á. Elbert [2] which asserts that $j_{\nu k}$ is a (strictly) concave function of ν for $-k < \nu < \infty$. The graph of such a function lies above each of its chords. Considering the chord joining $(0, j_{0k})$ and (N, j_{Nk}) we get

$$j_{\nu k} > j_{0k} + \nu(j_{Nk} - j_{0k})/N, \qquad 0 < \nu < N. \tag{2.1}$$

If we take $N = \frac{1}{2}$, this gives

$$j_{\nu k} > j_{0k} + 2\nu(k\pi - j_{0k}), \qquad 0 < \nu < \frac{1}{2}, \tag{2.2}$$

* Received by the editors June 22, 1981. This research was supported by the Consiglio Nazionale delle Ricerche of Italy, and by the Natural Sciences and Engineering Research Council of Canada.

† Istituto di Calcoli Numerici, Via Carlo Alberto 10, 10123 Torino, Italy.

‡ Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3.

which improves the lower bound

$$j_{\nu k} \geq \frac{\nu\pi}{2} + \left(k - \frac{1}{4}\right)\pi, \qquad 0 \leq \nu \leq \frac{1}{2}, \tag{2.3}$$

given by H. W. Hethcote [3, p. 73]. In fact, since it becomes exact at 0 and $\frac{1}{2}$, (2.2) gives the best lower bound for $j_{\nu k}$ which is *linear* on $(0, \frac{1}{2})$.

Letting $N \to \infty$ in (2.1) and taking account of (1.2), we get

$$j_{\nu k} \geq j_{0k} + \nu, \qquad 0 < \nu < \infty.$$

But the function

$$j_{\nu k} - j_{0k} - \nu$$

is concave and nonnegative on $(0, \infty)$ and hence cannot vanish there unless it is identically zero. Thus we have, in fact, for $k = 1, 2, \cdots$,

$$j_{\nu k} > j_{0k} + \nu, \qquad 0 < \nu < \infty. \tag{2.4}$$

This is an improvement on the inequality

$$j_{\nu k} > \left(j_{0k}^2 + \nu^2\right)^{1/2}, \qquad 0 < \nu < \infty,$$

found by R. C. McCann [6]. Recently, McCann [7] used a different method to prove (2.4) in the case $k = 1$. We remark that the asymptotic formula (1.2) shows that (2.4) is the best lower bound which is linear in ν and valid on $(0, \infty)$.

Next, we consider the chord joining the points $(-k, 0)$ and $(0, j_{0k})$ on the graph of $j_{\nu k}$ as a function of ν. The concavity of the graph gives

$$j_{\nu k} > j_{0k} + \left(\frac{\nu}{k}\right) j_{0k}, \qquad -k < \nu < 0. \tag{2.5}$$

Finally, the graph of the function $j_{\nu k}$ lies above the chord joining $(-\frac{1}{2}, (k - \frac{1}{2})\pi)$ and $(\frac{1}{2}, k\pi)$ leading to

$$j_{\nu k} > \left(k + \frac{\nu}{2} - \frac{1}{4}\right)\pi, \qquad -\frac{1}{2} < \nu < \frac{1}{2},$$

a result given in [4, p. 219] which extends the range of validity of (2.3).

The interval of validity of (2.4) cannot be extended to negative values of ν, but one can get linear lower bounds for various intervals (α, ∞), $-k < \alpha < 0$. The best linear lower bound for the interval $-k < \nu < \infty$ is easily seen to be

$$j_{\nu k} > \nu + k, \qquad -k < \nu < \infty.$$

3. Upper bounds for $j_{\nu k}$. Here we need the formula [11, p. 508]

$$\frac{dj_{\nu k}}{d\nu} = \frac{\pi}{2} j_{\nu k} \left[Y_\nu(z) \frac{\partial J_\nu(z)}{\partial \nu} - J_\nu(z) \frac{\partial Y_\nu(z)}{\partial \nu} \right]_{z = j_{\nu k}}. \tag{3.1}$$

But [8, p. 243]

$$\left[\frac{\partial J_\nu(z)}{\partial \nu}\right]_{\nu = 0} = \frac{\pi}{2} Y_0(z) \tag{3.2}$$

and [11, p. 76]

$$-Y_0(j_{0k}) = 2\left[\pi j_{0k} J_0'(j_{0k})\right]^{-1} = -2\left[\pi j_{0k} J_1(j_{0k})\right]^{-1}. \tag{3.3}$$

Thus

$$\left[\frac{dj_{\nu k}}{d\nu}\right]_{\nu=0}=\left[j_{0k}J_1^2(j_{0k})\right]^{-1}. \tag{3.4}$$

From Elbert's concavity result [2] we see that the graph of $j_{\nu k}$ lies below its tangent at $(0,j_{0k})$. This leads to the inequality

$$j_{\nu k}\leq j_{0k}+\nu\left[\frac{dj_{\nu k}}{d\nu}\right]_{\nu=0}, \qquad -k<\nu<\infty,$$

with equality only for $\nu=0$. In view of (3.4), this gives

$$j_{\nu k}<j_{0k}+\nu\left[j_{0k}J_1^2(j_{0k})\right]^{-1}, \qquad -k<\nu<\infty, \qquad \nu\neq 0. \tag{3.5}$$

Finally, the concavity of $j_{\nu k}$ shows that the function

$$F(\nu)=j_{0k}+\nu\left[j_{0k}J_1^2(j_{0k})\right]^{-1}-j_{\nu k}$$

increases with ν, $0<\nu<\infty$. In particular, we have

$$F(\nu)>F\left(\frac{1}{2}\right), \qquad \nu>\frac{1}{2},$$

leading to

$$j_{\nu k}<\left(\nu-\frac{1}{2}\right)\left[j_{0k}J_1^2(j_{0k})\right]^{-1}+k\pi, \qquad \nu>\frac{1}{2}. \tag{3.6}$$

It is clear that the concavity of $j_{\nu k}$ can be used to find many other inequalities. We have given only what appear to be the simplest ones.

In [5] it was shown that $j_{\nu k}^2$ is *convex* on a certain interval $\nu_k<\nu<\infty$ and conjectured that it is in fact convex on $0<\nu<\infty$. If this is indeed the case it would imply further inequalities for $j_{\nu k}$. Thus, for example one is led to *conjecture* that

$$j_{\nu k}^2<j_{0k}^2+2\nu\left[k^2\pi^2-j_{0k}^2\right], \qquad 0<\nu<\frac{1}{2}. \tag{3.7}$$

4. An approximation for zeros of Bessel functions of small order. Our main theorem here applies to zeros of $C_\nu(x)$. It is specialized to the zeros of $J_\nu(x)$ and $Y_\nu(x)$ in Corollaries 4.1 and 4.2, and numerical results for $j_{\nu 1}$ are given in §5.

THEOREM 4.1. *Let*

$$M(x)=\frac{\pi^2}{8}\left\{J_0^2(x)+Y_0^2(x)\right\}. \tag{4.1}$$

Then, for each fixed $k=1,2,\cdots$,

$$c_{\nu k}=c_{0k}+a_{1k}\nu+a_{2k}\nu^2+O(\nu^3), \qquad \nu\to 0, \tag{4.2}$$

where

$$a_{1k}=2c_{0k}M(c_{0k}) \tag{4.3}$$

and

$$a_{2k}=2c_{0k}M(c_{0k})\left[M(c_{0k})+c_{0k}M'(c_{0k})\right] \tag{4.4}$$

$$-\frac{\pi}{4}c_{0k}\left[J_\nu(x)\frac{\partial^2 Y_\nu(x)}{\partial\nu^2}-Y_\nu(x)\frac{\partial^2 J_\nu(x)}{\partial\nu^2}\right]_{\nu=0,x=c_{0k}}.$$

COROLLARY 4.1. *For each fixed* $k=1,2,\cdots$,

$$j_{\nu k}=j_{0k}+\alpha_{1k}\nu+\alpha_{2k}\nu^2+O(\nu^3), \qquad \nu\to 0,$$

where

$$\alpha_{1k}=\left[j_{0k}J_1^2(j_{0k})\right]^{-1}$$

and

$$\alpha_{2k}=\left[2j_{0k}^3J_1^4(j_{0k})\right]^{-1}-\frac{\pi}{2}\left[j_{0k}J_1^3(j_{0k})\right]^{-1}Y_1(j_{0k})$$
$$+\left[2J_1(j_{0k})\right]^{-1}\left[\frac{\partial^2 J_\nu(x)}{\partial\nu^2}\right]_{\nu=0,x=j_{0k}}.$$

COROLLARY 4.2. *Let* $y_{\nu k}$ *be the value taken on by* $c_{\nu k}$ *when* $\alpha=\pi/2$. (*Thus, for* $\nu\geq 0$, $y_{\nu k}$ *is the kth positive zero of* $Y_\nu(x)$.) *Then, for each* $k=1,2,\cdots$,

$$y_{\nu k}=y_{0k}+\beta_{1k}\nu+\beta_{2k}\nu^2+O(\nu^3), \qquad \nu\to 0,$$

where

$$\beta_{1k}=\frac{\pi^2}{4}y_{0k}J_0^2(y_{0k})$$

and

$$\beta_{2k}=\frac{\pi^4}{32}y_{0k}J_0^4(y_{0k})-\frac{\pi^4}{16}y_{0k}^2J_0^3(y_{0k})J_1(y_{0k})-\frac{\pi}{4}y_{0k}J_0(y_{0k})\left[\frac{\partial^2 Y_\nu(x)}{\partial\nu^2}\right]_{\nu=0,x=y_{0k}}.$$

Proof of Theorem 4.1. Nicholson's formula [11, p. 444], with $\nu=0$, may be written in view of (4.1), in the form

$$M(x)=\int_0^\infty K_0(2x\sinh t)\,dt, \qquad x>0. \tag{4.5}$$

We have [11, p. 446]

$$M(x)+xM'(x)=\int_0^\infty K_0(2x\sinh t)\tanh^2 t\,dt, \qquad x>0. \tag{4.6}$$

We also have [11, p. 444]

$$J_\nu(x)\frac{\partial Y_\nu(x)}{\partial\nu}-Y_\nu(x)\frac{\partial J_\nu(x)}{\partial\nu}$$
$$=-\frac{4}{\pi}\int_0^\infty K_0(2x\sinh t)e^{-2\nu t}\,dt, \qquad x>0, \tag{4.7}$$

and [11, p. 508]

$$\frac{dc_{\nu k}}{d\nu}=2c_{\nu k}\int_0^\infty K_0(2c_{\nu k}\sinh t)e^{-2\nu t}\,dt. \tag{4.8}$$

Differentiation of (4.8) and integration by parts lead to

$$\left[\frac{d^2c_{\nu k}}{d\nu^2}\right]_{\nu=0}=2\left[\frac{dc_{\nu k}}{d\nu}\right]_{\nu=0}\int_0^\infty K_0(2c_{0k}\sinh t)\tanh^2 t\,dt$$
$$-4c_{0k}\int_0^\infty tK_0(2c_{0k}\sinh t)\,dt. \tag{4.9}$$

Elbert [2, (16), p. 87] has this in the case $c_{\nu k}=j_{\nu k}$; it is clear that his method applies in the more general case. It follows from (4.7) that

$$(4.10)\quad J_\nu(x)\frac{\partial^2 Y_\nu(x)}{\partial \nu^2}-Y_\nu(x)\frac{\partial^2 J_\nu(x)}{\partial \nu^2}=\frac{8}{\pi}\int_0^\infty tK_0(2x\sinh t)e^{-2\nu t}\,dt, \qquad x>0.$$

Now since $c_{\nu k}$ is analytic in ν [11, p. 507], the expansion (4.2) holds with

$$a_{1k}=\left[\frac{dc_{\nu k}}{d\nu}\right]_{\nu=0}=2c_{0k}\int_0^\infty K_0(2c_{0k}\sinh t)\,dt,$$

using (4.8), and this is easily seen to give (4.3), on using (4.5). Next,

$$a_{2k}=\frac{1}{2}\left[\frac{d^2c_{\nu k}}{d\nu^2}\right]_{\nu=0}.$$

Using (4.9) and the result just found for a_{1k} gives

$$a_{2k}=2c_{0k}M(c_{0k})\int_0^\infty K_0(2c_{0k}\sinh t)\tanh^2 t\,dt$$
$$-2c_{0k}\int_0^\infty tK_0(2c_{0k}\sinh t)\,dt.$$

We now use (4.6) and (4.10) to get (4.4).

The corollaries follow easily from the theorem on using (4.1), the Wronskian formula for $J_\nu(x)$ and $Y_\nu(x)$ [11, p. 76] and some simple recurrence relations.

An alternative proof of Theorem 4.1 (at least in the case of the zeros of $J_\nu(x)$) may be based on the results of F. G. Tricomi [9] on approximating zeros of functions for which asymptotic expansions are known.

5. Numerical results for $j_{\nu 1}$. Corollary 4.1 shows that

$$j_{\nu 1}=j_{01}+\alpha_{11}\nu+\alpha_{21}\nu^2+O(\nu^3), \qquad \nu\to 0,$$

where

$$\alpha_{11}=\left[j_{01}J_1^2(j_{01})\right]^{-1}$$

and

$$\alpha_{21}=\left[2j_{01}^3J_1^4(j_{01})\right]^{-1}-\frac{\pi}{2}\left[j_{01}J_1^3(j_{01})\right]^{-1}Y_1(j_{01})+\left[2J_1(j_{01})\right]^{-1}\left[\frac{\partial^2 J_\nu(x)}{\partial\nu^2}\right]_{\nu=0,x=j_{01}}.$$

The values of j_{01} and $J_1(j_{01})=-J_0'(j_{01})$ needed to evaluate α_{11} are found in [1, p. 409]. For α_{21} we need in addition to evaluate $Y_1(j_{01})$ and

$$(5.1)\qquad \left[\frac{\partial^2 J_\nu(x)}{\partial\nu^2}\right]_{\nu=0,x=j_{01}}.$$

For the first of these we use the series expansion of $Y_1(x)$, while for (5.1) we use [11, p. 61]

$$(5.2)\quad \frac{\partial J_\nu(x)}{\partial\nu}=\sum_{k=0}^\infty(-1)^k\left(\frac{x}{2}\right)^{\nu+2k}\left[k!\Gamma(\nu+k+1)\right]^{-1}\cdot\left[\log\frac{x}{2}-\psi(\nu+k+1)\right]$$

and its consequence

(5.3)

$$\frac{\partial^2 J_\nu(x)}{\partial \nu^2}=2\log\frac{x}{2}\,\frac{\partial J_\nu(x)}{\partial \nu}-J_\nu(x)\left(\log\frac{x}{2}\right)^2 + \sum_{k=0}^{\infty}(-1)^k\left(\frac{x}{2}\right)^{\nu+2k}[k!\Gamma(\nu+k+1)]^{-1}[\psi^2(\nu+k+1)-\psi'(\nu+k+1)].$$

Using (3.2) leads to

(5.4)
$$\left[\frac{\partial^2 J_\nu(x)}{\partial \nu^2}\right]_{\nu=0,x=j_{\nu 1}}=2\log\left(\frac{j_{01}}{2}\right)[j_{01}J_1(j_{01})]^{-1} + \sum_{k=0}^{\infty}(-1)^k\left(\frac{j_{01}}{2}\right)^{2k}(k!)^{-2}[\psi^2(k+1)-\psi'(k+1)].$$

Here $\psi(x)=\Gamma'(x)/\Gamma(x)$ so that to compute the sum of the series in (5.4) we may use [1, pp. 258–260]

$$\psi(n+1)=\psi(n)+\frac{1}{n}, \qquad \psi(1)=-\gamma,$$

$$\psi'(n+1)=\psi'(n)-\frac{1}{n^2}, \qquad \psi'(1)=\frac{\pi^2}{6}.$$

By means of straightforward calculations we are led to the numerical values in (1.1).

Acknowledgments. We thank Professor L. Gatteschi for his encouragement and for useful discussions. The second author is grateful to Professor L. Lorch for a discussion of inequality (2.4) and for the hospitality extended by the Department of Mathematics, University of Dundee, where he was a visitor when this paper was written.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Applied Mathematics Series, 55, National Bureau of Standards, Washington, DC, 1964.

[2] Á. Elbert, *Concavity of the zeros of Bessel functions*, Studia Sci. Math. Hungar., 12 (1977), pp. 81–88.

[3] H. W. Hethcote, *Bounds for zeros of some special functions*, Proc. Amer. Math. Soc., 25 (1970), pp. 72–74.

[4] A. Laforgia, *Sugli zeri delle funzioni di Bessel*, Calcolo, 17 (1980), pp. 211–220.

[5] J. T. Lewis and M. E. Muldoon, *Monotonicity and convexity properties of zeros of Bessel functions*, this Journal, 8 (1977), pp. 171–178.

[6] R. C. McCann, *Lower bounds for zeros of Bessel functions*, Proc. Amer. Math. Soc., 64 (1977), pp. 101–103.

[7] ———, *Monotonicity properties of zeros of Bessel functions*, Abstracts Amer. Math Soc., 2 (1981), p. 105.

[8] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York and London, 1974.

[9] F. G. Tricomi, *Sugli zeri delle funzioni di cui si conosce una rappresentazione asintotica*, Ann. Mat. Pura Appl. (4), 26 (1947), pp. 283–300.

[10] ———, *Sulle funzioni di Bessel di ordine e argomento pressoché uguali*, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 83 (1949), pp. 3–20.

[11] G. N. Watson, *A Treatise on the Theory of Bessel functions*, 2nd ed., Cambridge Univ. Press, London, 1944.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0013 $01.25/0

TOTAL POSITIVITY OF MEAN VALUES AND HYPERGEOMETRIC FUNCTIONS*

B. C. CARLSON† AND JOHN L. GUSTAFSON†

Abstract. The weighted power mean of two positive variables is strictly totally positive (STP) if its order t satisfies $-\infty<t<0$, and its reciprocal is STP if $0<t<\infty$. The reciprocals of the logarithmic mean, Gauss's arithmetic-geometric mean, and the Schwab–Borchardt mean are STP. The hypergeometric R-function $R_{-\alpha}(\beta,\beta';x,y)$, $x,y>0$, which is equivalent to ${}_2F_1$ with argument $1-x/y$, is STP if $\alpha,\beta\beta'$, and $\beta+\beta'-\alpha$ are positive. With weaker restrictions this function is represented in a new way as a convolution. Higher order positivity is discussed for some other hypergeometric functions, including incomplete elliptic integrals.

1. Introduction. A real-valued function $f(x,y)$ of two real variables is said to be strictly totally positive (STP) on its domain of definition if every $n\times n$ determinant with elements $f(x_i,y_j)$, where $x_1<x_2<\cdots<x_n$ and $y_1<y_2<\cdots<y_n$, is strictly positive for every $n=1,2,\cdots$. If the determinants are strictly positive for $n=1,2,\cdots,r$, then f is said to be strictly positive of order r (SP_r). The principal reference for the subject is Karlin [6], who writes STP_r in place of SP_r and sometimes STP_∞ for STP. Many applications to statistics, mechanics, and differential equations arise from the circumstance that a totally positive function is the kernel of a variation-diminishing transform.

We refer to [6] or [7, Chap. 18] for more precise statements and proofs of several basic facts:

(1.1) e^{xy} is STP for x,y real [6, pp. 15–16].

(1.2) If both g and h are strictly increasing functions, or if both are strictly decreasing, and if $F(x,y)=f(g(x),h(y))$, then F is SP_r if f is SP_r [6, p. 18].

(1.3) If g and h are strictly positive functions, and if $F(x,y)=g(x)f(x,y)h(y)$, then F is SP_r if f is SP_r [6, p. 18].

(1.4) If $f(x,y)=\int_Z g(x,z)h(z,y)\,d\sigma(z)$, where σ is a positive σ-finite measure on Z and the integral converges absolutely, then f is SP_r on $X\times Y$ if g is SP_r on $X\times Z$ and h is SP_r on $Z\times Y$ [6, pp. 16–17].

To these four rules we add two more:

(1.5) If (1.4) is modified so that either

$$\frac{1}{f(x,y)}=\int_Z \frac{h(z,y)}{g(x,z)}\,d\sigma(z) \quad \text{or} \quad f(x,y)=\int_Z \frac{d\sigma(z)}{g(x,z)h(z,y)},$$

then f is SP_2 if g and h are SP_2. This follows from [6, Eq. (2.5)] and the observation that $a_{11}, a_{12}, a_{21}, a_{22}>0$ implies that the 2×2 determinant with elements a_{ij} is strictly positive if and only if the 2×2 determinant with elements $1/a_{ij}$ is strictly negative.

(1.6) If $a>0$ then $(x+y)^{-a}$ is STP for $x,y>0$.

Apparently (1.6) is new except for the case $a=1$ [6, pp. 149–150], which dates back to Cauchy and demonstrates that all minors of the Hilbert matrix are positive. The

*Received by the editors October 19, 1981.

†Ames Laboratory and Department of Mathematics, Iowa State University, Ames, Iowa 50011. Ames Laboratory is operated for the U. S. Department of Energy by Iowa State University under contract W-7405-Eng-82. This research was supported by the Director of Energy Research, Office of Basic Energy Sciences.

proof of the general case follows from the integral representation of the gamma function [2, Ex. 3.2-3],

$$(x+y)^{-a}\Gamma(a)=\int_0^\infty t^{a-1}e^{-(x+y)t}\,dt,$$

$$(x+y)^{-a}=\int_{-\infty}^0 e^{xz}e^{zy}\,d\sigma(z),$$

where $d\sigma(z)=(-z)^{a-1}\,dz/\Gamma(a)$. The proof is completed by using (1.1) and (1.4).

2. Power means. The weighted power mean [4, p. 13] of order t is defined by

$$M_t(x,y)=\left[wx^t+(1-w)y^t\right]^{1/t}, \qquad t\neq 0, \tag{2.1}$$

where $x,y>0$ and $0<w<1$.

THEOREM 2.1. *If $0<t<\infty$ then $1/M_t(x,y)$ is STP for $x,y>0$. If $-\infty<t<0$ then $M_t(x,y)$ is STP for $x,y>0$.*

Proof. It follows from (1.6) and (1.2) that $[wx^t+(1-w)y^t]^{-a}$ is STP if $a>0$ and $t\neq 0$. Assuming $0<t<\infty$ and putting $a=1/t$, we conclude that $1/M_t(x,y)$ is STP. If $-\infty<t<0$ we put $a=-1/t$.

Note that the geometric mean, $M_0(x,y)=x^w y^{1-w}$, is not STP because the rows of the relevant determinants are proportional. The possibility of proportional rows likewise keeps M_∞ and $M_{-\infty}$ [4, p. 15] from being STP, although the determinants are nonnegative.

If $a>0$ and $c\geq 0$, $(x+y+c)^{-a}$ is STP for $x,y>0$ by (1.6) and (1.2). Hence the weighted power mean of several variables, $[\Sigma w_i x_i^t]^{1/t}$, has the positivity properties of Theorem 2.1 in any two of the variables if the others are held fixed.

3. Iterative means. If $x,y>0$ let $x_0=x$ and $y_0=y$ and consider three separate iterative processes in which x_n and y_n approach a common limit as $n\to\infty$:

$$x_{n+1}=\frac{1}{2}x_n+\frac{1}{2}(x_n y_n)^{1/2}, \qquad y_{n+1}=\frac{1}{2}y_n+\frac{1}{2}(x_n y_n)^{1/2}, \qquad x_n, y_n\to L(x,y), \tag{3.1}$$

$$x_{n+1}=\frac{1}{2}(x_n+y_n), \qquad y_{n+1}=(x_n y_n)^{1/2}, \qquad x_n, y_n\to M(x,y), \tag{3.2}$$

$$x_{n+1}=\frac{1}{2}(x_n+y_n), \qquad y_{n+1}=(x_{n+1} y_n)^{1/2}, \qquad x_n, y_n\to S(x,y). \tag{3.3}$$

Here L is the logarithmic mean, M is Gauss's arithmetic-geometric mean, and S is the Schwab–Borchardt mean[1]. The reciprocal of each has an integral representation [1]:

$$\frac{1}{L(x,y)}=R_{-1}(1,1;x,y)=\frac{\ln x-\ln y}{x-y}, \tag{3.4}$$

$$\frac{1}{M(x,y)}=R_{-1/2}\left(\frac{1}{2},\frac{1}{2};x^2,y^2\right), \tag{3.5}$$

$$\frac{1}{S(x,y)}=R_{-1/2}\left(\frac{1}{2},1;x^2,y^2\right)=\begin{cases}(y^2-x^2)^{-1/2}\arccos(x/y), & x<y,\\ (x^2-y^2)^{-1/2}\operatorname{arccosh}(x/y), & x>y,\end{cases} \tag{3.6}$$

[1] The iterative process converging to S was proposed but not published by Gauss in 1800 (for more details see [1]). Schwab [9, pp. 103–107] published it in 1813 and Borchardt in 1880. We thank Professor I. J. Schoenberg for reference [9].

where

$$R_{-\alpha}(\beta,\beta';x,y)=\int_0^\infty (x+z)^{-\beta}(z+y)^{-\beta'}d\sigma(z),$$

$$d\sigma(z)=\frac{\Gamma(\beta+\beta')}{\Gamma(\alpha)\Gamma(\beta+\beta'-\alpha)}z^{\beta+\beta'-\alpha-1}dz, \qquad 0<\alpha<\beta+\beta'. \tag{3.7}$$

It follows from (1.4) and (1.6) that $R_{-\alpha}(\beta,\beta';x,y)$ is STP for $x,y>0$ provided $\beta,\beta'>0$ and $0<\alpha<\beta+\beta'$. Use of (1.2) completes the proof of the following theorem:

THEOREM 3.1. *The reciprocal means* $1/L(x,y)$, $1/M(x,y)$, *and* $1/S(x,y)$ *are STP for* $x,y>0$.

The means M and S are the best-known members of a family of twelve iterative means $L_{ij}(x,y)$ constructed by letting

$$x_{n+1}=f_i(x_n,y_n), \qquad y_{n+1}=f_j(x_n,y_n), \qquad i\neq j, \tag{3.8}$$

where

$$f_1(x,y)=\frac{1}{2}(x+y), \qquad f_2(x,y)=(xy)^{1/2},$$

$$f_3(x,y)=\left(x\frac{x+y}{2}\right)^{1/2}, \qquad f_4(x,y)=\left(y\frac{x+y}{2}\right)^{1/2}. \tag{3.9}$$

For each of the twelve choices of i and j, $i\neq j$, the common limit of x_n and y_n as $n\to\infty$ is $L_{ij}(x,y)$. For example the Schwab–Borchardt mean S is L_{14}. In each case a suitable negative power ($-1/2$ or -1 or -2) of L_{ij} (see [1]) is an R-function (3.7) with α,β,β' such that it is STP. The mean L also is essentially a member of this family, as one sees by replacing each variable in (3.1) by its square.

4. Hypergeometric functions. The R-function (3.7) is a homogeneous variant of Gauss's hypergeometric function [2, §5.9]:

$$R_{-\alpha}(\beta,\beta';x,y)=y^{-\alpha}{}_2F_1\left(\alpha,\beta;\beta+\beta';1-\frac{x}{y}\right). \tag{4.1}$$

If b is a k-tuple of real numbers and x a k-tuple of positive numbers, an extension of (3.7) to several variables is [2, (6.8-6)]

$$R_{-a}(b,x)=\int_0^\infty \prod_{i=1}^k (x_i+z)^{-b_i}d\sigma(z),$$

$$d\sigma(z)=\frac{\Gamma(a+a')}{\Gamma(a)\Gamma(a')}z^{a'-1}dz, \qquad a'=\sum_{i=1}^k b_i-a, \qquad a>0, \quad a'>0. \tag{4.2}$$

The R-function has other representations that define it when a and a' are not positive.

THEOREM 4.1. *Let* $a,a',b_1,\cdots,b_k$ *be real numbers and assume* $a+a'=\sum_{i=1}^k b_i$ *and* $aa'b_1\cdots b_k\neq 0$. *Let* $x_i>0$, $i=1,\cdots,k$. *For some* i *and* j *consider* $R_{-a}(b,x)$ *as a function of* x_i *and* x_j, *all other components of* x *being fixed; i.e., define* $f(x_i,x_j)=[(x_i,x_j)\mapsto R_{-a}(b,x)]$. *If* $k\geq 2$ *and* $a,a',b_i,b_j>0$, *then* f *is STP. If* $k=2$ *and exactly one of* a,a',b_1,b_2 *is negative, then* $1/f$ *is* SP_2. *If* $k>2$ *and* $a,a'>0$, *then* $1/f$ *is* SP_2 *if* $b_ib_j<0$ *while* f *is* SP_2 *if* $b_i<0$ *and* $b_j<0$.

Proof. In those parts of the theorem which assume $a, a' > 0$, we may use (4.2) and define a sigma-finite measure

$$d\sigma_1(z) = \prod_{m \neq i,j} (x_m + z)^{-b_m} d\sigma(z).$$

If b_i, $b_j > 0$ then (1.6) and (1.4) imply that f is STP. If $b_i < 0$ then $(x_i + z)^{-b_i}$ is the reciprocal of a function that is STP and therefore SP_2. Hence the last sentence of the theorem follows from (1.5), as does the next to last sentence in case exactly one of b_1 and b_2 is negative. The remaining case, when exactly one of a and a' is negative while b_1 and b_2 are positive, follows from [2, (5.9-20)] and (1.3).

Theorem 4.1 has interesting applications to elliptic integrals. For example, the perimeter of an ellipse [2, (9.4-5)] with semiaxes α and β is $P(\alpha, \beta) = 2\pi R_{1/2}(\frac{1}{2}, \frac{1}{2}; \alpha^2, \beta^2)$, and hence $1/P(\alpha, \beta)$ is SP_2 for $\alpha, \beta > 0$. The symmetric incomplete integrals of the first and third kinds [3],

$$R_F(x,y,z) = R_{-1/2}\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; x, y, z\right), \quad R_J(x,y,z,p) = R_{-3/2}\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1; x, y, z, p\right),$$

where $x, y, z, p > 0$, are STP in any two variables when the others are fixed. We may choose $z = 1$ by homogeneity and tabulate $R_F(x, y, 1)$ with rows and columns of the table labeled by increasing values of x and y, respectively. If the table is regarded as a matrix, all its minors are strictly positive. Similar remarks apply to the integral of the second kind, $R_D(x,y,z) = R_J(x,y,z,z) = R_{-3/2}(\frac{1}{2}, \frac{1}{2}, \frac{3}{2}; x, y, z)$.

Theorem 4.1 implies that $1/R_t(\beta, \beta'; x, y)$ is SP_2 for $x, y > 0$ provided $\beta, \beta' > 0$ and either $t > 0$ or $t < -\beta - \beta'$. We ask now whether the SP_2 property can be strengthened to STP or at least SP_r for some $r > 2$. Because of [2, (5.9-21)] and (1.3), $1/R_t(\beta, \beta'; x, y)$ is SP_r if and only if $1/R_{-\beta-\beta'-t}(\beta, \beta'; x, y)$ is SP_r. Hence it suffices to consider the case $t > 0$.

If $\beta, \beta' > 0$, it is not hard to show that $1/R_t(\beta, \beta'; x, y)$ is STP for $x, y > 0$ in certain special and limiting cases. If $t = 1$ we use [2, (6.2-2)]. For any $t > 0$, as $\beta + \beta'$ tends to 0 or ∞ with β/β' fixed, we use [2, (6.2-17), (6.2-18)]. (The cited equations are valid also for nonintegral n.) Some additional special cases in which $1/R_t$ is STP if $t > 0$ will be exhibited in §5.

Nevertheless, a numerical example shows that $1/R_2(\frac{1}{2}, \frac{1}{2}; x, y)$ is not SP_3. If $(x_1, x_2, x_3) = (1, 2, 3)$ and $(y_1, y_2, y_3) = (100, 200, 300)$, the 3×3 determinant with elements $1/R_2(\frac{1}{2}, \frac{1}{2}; x_i, y_j)$ has the value -1.7×10^{-20}. More generally a complicated algebraic expression for the 3×3 determinant with elements $1/R_2(\beta, \beta'; x_i, y_j)$ shows that the determinant will be negative for fixed positive $\beta < 1$ if x_3/y_1 (or y_3/x_1) is sufficiently small.

We conclude that if $t > 0$ or $t < -\beta - \beta'$, then $1/R_t$ is sometimes STP and sometimes not even SP_3 but always SP_2 if $\beta, \beta' > 0$. Some further examples in which it is or is not STP will be discussed in the next section by using the properties of Pólya frequency functions.

Since the weighted power mean (2.1) of order t is the limit as $c \to 0+$ of the hypergeometric mean $[R_t(cw, c - cw; x, y)]^{1/t}$, it is natural to ask whether the reciprocal of the latter is STP if $c > 0$ and $t > -c$. In general it is not. For instance, if $(x_1, x_2, x_3) = (1, 2, 3)$ and $(y_1, y_2, y_3) = (100, 200, 300)$, the 3×3 determinant with elements $1/[R_2(\frac{1}{2}, \frac{1}{2}; x_i, y_j)]^{1/2}$ has the value -8.1×10^{-15}.

5. Pólya frequency functions. A measurable real-valued function f defined on the real line is called a strict Pólya frequency function (SPF) if $f(x - y)$ is STP. (Some authors require f to be integrable, but if f is SPF then $e^{cx} f(x)$ is integrable for suitable

real c [8, p. 341].) If $f(x-y)$ is SP_r then f is called SPF_r. A function is SPF_2 if and only if it is strictly log-concave on the real line [8, p. 337].

For example, if $\beta, \beta' > 0$ and $0 < \alpha < \beta + \beta'$, then $R_{-\alpha}(\beta, \beta'; e^{2x}, e^{2y})$ is STP for real x and y by Theorem 4.1 and (1.2). Since $R_{-\alpha}$ is homogeneous of degree $-\alpha$, we have

$$R_{-\alpha}(\beta, \beta'; e^{2x}, e^{2y}) = e^{-\alpha x} e^{-\alpha y} R_{-\alpha}(\beta, \beta'; e^{x-y}, e^{y-x}).$$

It follows by (1.3) that $R_{-\alpha}(\beta, \beta'; e^{x}, e^{-x})$ is SPF.

For another example, the Gegenbauer polynomial [2, (6.7-21)] of degree n is

$$C_n^{\nu}(\cosh x) = \frac{\Gamma(2\nu+n)}{\Gamma(2\nu)\Gamma(n+1)} R_n(\nu, \nu; e^{x}, e^{-x}). \tag{5.1}$$

If $\nu > 0$ and $n = 1, 2, 3, \cdots$, it follows from Theorem 4.1 that $1/C_n^{\nu}(\cosh x)$ is SPF_2 and $C_n^{\nu}(\cosh x)$ is strictly log-convex. The same is true for the Gegenbauer function defined by (5.1) with any real $n > 0$ and $\nu > 0$.

To see whether $1/C_n^{\nu}$ is SPF, we shall use a theorem of Schoenberg [8, p. 349] with strictness conditions added by Karlin [6, p. 357]. Only an abridged version of the theorem will be needed. A measurable real-valued function f defined on the real line is SPF if its bilateral Laplace transform exists in an open strip containing the imaginary axis and has the form

$$\int_{-\infty}^{\infty} e^{-sx} f(x)\, dx = \frac{1}{\varphi(s)}, \qquad \varphi(s) = Ce^{\delta s} \prod_{i=1}^{\infty} (1 + a_i s) e^{-a_i s}, \tag{5.2}$$

where $C > 0$, the a_i and δ are real, Σa_i^2 converges, and $\Sigma |a_i|$ diverges. Conversely, f is not SPF unless the reciprocal of its bilateral Laplace transform is entire.

For example, if $\beta, \beta' > 0$, $-\alpha < \mathrm{Re}\, s < \alpha$, and $\alpha - 2\beta < \mathrm{Re}\, s < 2\beta' - \alpha$, then

$$\int_{-\infty}^{\infty} e^{-sx} R_{-\alpha}(\beta, \beta'; e^{x}, e^{-x})\, dx = \frac{\Gamma(\beta+\beta')\Gamma\left(\frac{\alpha+s}{2}\right)\Gamma\left(\frac{\alpha-s}{2}\right)\Gamma\left(\frac{2\beta-\alpha+s}{2}\right)\Gamma\left(\frac{2\beta'-\alpha-s}{2}\right)}{2\Gamma(\beta)\Gamma(\beta')\Gamma(\alpha)\Gamma(\beta+\beta'-\alpha)}, \tag{5.3}$$

as one finds by taking e^{-x} as a new integration variable to obtain a Mellin transform, substituting (3.7), and changing the order of integration. The representation of Γ by an infinite product shows that (5.3) has the form (5.2). This was expected, since the conditions of validity imply $0 < \alpha < \beta + \beta'$.

Since the product of the Laplace transforms of two functions is the transform of their convolution, (5.3) suggests a new way of writing the hypergeometric function (4.1) as a convolution:

$$R_{-\alpha}(\beta, \beta'; e^{x}, e^{-x}) = \frac{2^{1-\beta-\beta'}}{B(\beta, \beta')} \int_{-\infty}^{\infty} \mathrm{sech}^{\alpha}(x-t) e^{(\beta'-\beta)t} (\mathrm{sech}\, t)^{\beta+\beta'-\alpha} dt, \tag{5.4}$$

where $|\mathrm{Im}\, x| < \pi/2$, $\mathrm{Re}\, \beta > 0$, and $\mathrm{Re}\, \beta' > 0$. These conditions of validity can be verified by putting $e^{2t} = (1-u)/u$ to obtain Euler's representation. Equation (5.4) is particularly attractive if β and β' are equal, as they are for Legendre and Gegenbauer functions [2, §6.8].

We can now investigate further the higher order positivity of $1/R_t$, $t>0$. For example,

$$(5.5)\qquad \int_{-\infty}^{\infty}\frac{e^{-sx}\,dx}{R_t(1,1;e^x,e^{-x})}=\frac{\pi\sin\left(\frac{\pi}{t+1}\right)}{2\sin\left(\frac{\pi}{2}\frac{t+s}{t+1}\right)\sin\left(\frac{\pi}{2}\frac{t-s}{t+1}\right)},\qquad -t<\mathrm{Re}\,s<t.$$

This result follows from (5.3): observe that [2, Ex. 5.9-13]

$$\frac{1}{R_t(1,1;e^x,e^{-x})}=\frac{(t+1)\sinh x}{\sinh[(t+1)x]}=R_{-t/(t+1)}(1,1;e^y,e^{-y}),\qquad y=(t+1)x.$$

The representation of the sine function by an infinite product shows that (5.5) has the form (5.2). Hence $1/R_t(1,1;e^x,e^{-x})$, $t>0$, is SPF and $1/R_t(1,1;x,y)$, $t>0$, is STP for $x,y>0$.

Another example, in which the Laplace transform can be evaluated by using [2, Ex. 6.10-12, (4.2-4)] after taking e^{-x} as a new variable of integration, is

$$(5.6)\qquad \int_{-\infty}^{\infty}\frac{e^{-sx}\,dx}{R_t\left(\frac{1}{2}-t,\frac{1}{2}-t;e^x,e^{-x}\right)}=\frac{2^{2t}\Gamma(t+s)\Gamma(t-s)}{\Gamma(2t)},$$

where $-t<\mathrm{Re}\,s<t$ and $t\neq 1,2,3,\cdots$. Since this has the form (5.2), the condition $0<t\neq 1,2,3,\cdots$ ensures that $1/R_t(\frac{1}{2}-t,\frac{1}{2}-t;e^x,e^{-x})$ is SPF and $1/R_t(\frac{1}{2}-t,\frac{1}{2}-t;x,y)$ is STP for $x,y>0$. The same is true of $1/R_{t-1}(\frac{1}{2}-t,\frac{1}{2}-t;x,y)$ with the same conditions on t,x,y (see [2, Ex. 6.10-12]).

Despite the preceding special cases (as well as the cases mentioned near the end of §4) in which $1/R_t$, $t>0$, is STP, a final example suggests that this state of affairs may be the exception rather than the rule. If β,β' are real and $\beta\beta'(\beta+\beta'+1)>0$, then [2, (6.2-4)] yields

$$(5.7)\qquad \int_{-\infty}^{\infty}\frac{e^{-sx}\,dx}{R_2(\beta,\beta';e^x,e^{-x})}=\frac{\pi}{2}(\beta+\beta')(\tan\theta)\left[\frac{\beta(\beta+1)}{\beta'(\beta'+1)}\right]^{s/4}\frac{\sin(s\theta/2)}{\sin(s\pi/2)},$$

where $-2<\mathrm{Re}\,s<2$, $0<\theta<\pi/2$, and $\tan\theta=[(\beta+\beta'+1)/\beta\beta']^{1/2}$. If $\pi/\theta=3,4,5,\cdots$, all zeros of $\sin(s\theta/2)$ are cancelled by zeros of $\sin(s\pi/2)$. Then (5.7) has the form (5.2) and $1/R_2(\beta,\beta';e^x,e^{-x})$ is SPF. (The case $\beta=\beta'=1$ coincides with the case $t=2$ of (5.5).) In particular, by (5.1), $1/C_2^{\nu}(\cosh x)$ is SPF if $\nu/(\nu+1)=\cos(\pi/m)$, $m=3,4,5,\cdots$. On the other hand, if $0<\theta<\pi/2$ but θ does not have one of the listed values, the reciprocal of the Laplace transform is not entire and $1/R_2(\beta,\beta';x,y)$ is not STP. The numerical example in §4 shows that it is not always even SP_3.

Other interesting examples of sign regularity properties of hypergeometric functions are contained in [10].

REFERENCES

[1] B. C. CARLSON, *Algorithms involving arithmetic and geometric means*, Amer. Math. Monthly, 78 (1971), pp. 496–505.

[2] _____, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[3] _____, *Computing elliptic integrals by duplication*, Numer. Math., 33 (1979), pp. 1–16.

[4] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge Univ. Press, Cambridge, 1952.

[5] I. I. HIRSCHMAN AND D. V. WIDDER, *The Convolution Transform*, Princeton Univ. Press, Princeton, NJ, 1955.
[6] S. KARLIN, *Total Positivity*, Stanford Univ. Press, Stanford, CA, 1968.
[7] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[8] I. J. SCHOENBERG, *On Pólya frequency functions*, I. *The totally positive functions and their Laplace transforms*, J. d'Anal. Math., 1 (1951), pp. 331–374.
[9] J. SCHWAB, *Élémens de géométrie*, vol. 1, C.-J. Hissette, Nancy, 1813.
[10] S. KARLIN, *Sign regularity properties of classical orthogonal polynomials*, in Orthogonal Expansions and Their Continuous Analogues, D. T. Haimo, ed., Southern Illinois Univ. Press, Carbondale, 1968, pp. 55–74.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0014 $01.25/0

ON ELEMENTARY METHODS IN POSITIVITY THEORY*

J. GILLIS,† B. REZNICK‡ AND D. ZEILBERGER§

Abstract. We give a short proof of a result of Askey and Gasper [J. Analyse Math., 31 (1977), pp. 48–68] that $(1-x-y-z+4xyz)^{-\beta}$ has positive power series coefficients for $\beta \geq (\sqrt{17}-3)/2$. We also show how Ismail and Tamhankar's proof [SIAM J. Math. Anal., 10 (1979), pp. 478–485] that

$$(1-(1-\lambda)x-\lambda y-\lambda xz-(1-\lambda)yz+xyz)^{-\alpha} \qquad (0\leq\lambda\leq 1)$$

has positive power series coefficients for $\alpha=1$ implies Koornwinder's result that it does so for $\alpha\geq 1$.

1. Introduction. Given a multivariate polynomial $P(x_1,\cdots,x_n)$ and a real β, it is of interest to know whether $P^{-\beta}$ has only positive terms in its power series expansion. Szegö [6] proved that this was the case for $P=(1-x)(1-y)+(1-x)(1-z)+(1-y)\cdot(1-z)$ and $\beta\geq\frac{1}{2}$, and Askey and Gasper [2] established positivity for $P=1-x-y-z+4xyz$ and $\beta\geq(\sqrt{17}-3)/2$. A fascinating account of the history of these problems up to 1975 is given in Askey's monograph [1].

Koornwinder [5] used deep methods to establish the positivity of the coefficients of $[1-(1-\lambda)x-\lambda y-\lambda xz-(1-\lambda)yz+xyz]^{-\beta}$ for $0\leq\lambda\leq 1$, $\beta\geq 1$ and that of $[1-x-y-z-u+4(xyz+xyu+xzu+yzu)-16xyzu]^{-\beta}$. Later, Ismail and Tamhankar [4] (see also [3]) gave elementary proofs of Koornwinder's results in the special case $\beta=1$. In §2 we are going to show how Ismail and Tamhankar's results for $\beta=1$ imply Koornwinder's results for $\beta\geq 1$ and in §3 we give a short proof of Askey and Gasper's [2] result. Finally, in §4 we conjecture that for $n\geq 4$, $(1-(x_1+\cdots+x_n)+n!x_1\cdots x_n)^{-1}$ has positive coefficients.

2. Operations that preserve positivity of coefficients.

PROPOSITION 1. *Suppose that $a(x_1,\cdots,x_{n-1})$ and $b(x_1,\cdots,x_{n-1})$ are polynomials. If* (i) $(a-bx_n)^{-1}$ *has positive coefficients and* (ii) $a^{-\alpha}$ *has positive coefficients for all $\alpha>0$, then so does $(a-bx_n)^{-\beta}$ for all $\beta\geq 1$.*

Proof. By hypothesis $(a-bx_n)^{-1}=\Sigma(b^r/a^{r+1})x_n^r$ has positive coefficients, implying that for every r, b^r/a^{r+1} has positive coefficients. Since $(\beta)_r/r!=\beta(\beta+1)\cdots(\beta+r-1)/r!$ is positive and $a^{1-\beta}$ has positive coefficients, we see that

$$(a-bx_n)^{-\beta}=a^{1-\beta}\sum_{r=0}^{\infty}\frac{(\beta)_r}{r!}\frac{b^r}{a^{r+1}}$$

has positive coefficients. □

By taking $a=1-(1-\lambda)x-\lambda y$, $b=\lambda x+(1-\lambda)y-xy$ $(0\leq\lambda\leq 1)$ it follows that Ismail and Tamhankar's result that $[1-(1-\lambda)x-\lambda y-\lambda xz-(1-\lambda)yz+xyz]^{-\beta}$ $(0\leq\lambda\leq 1)$ has positive coefficients for $\beta=1$ implies Koornwinder's result that it does so for $\beta\geq 1$.

PROPOSITION 2. *If $[a(x,y)-b(x,y)z]^{-\alpha}$ and $[c(x,y)-d(x,y)z]^{-\alpha}$ have positive coefficients ($\alpha>0$) so also does $[a(x,y)c(z,u)-b(x,y)d(z,u)]^{-\alpha}$.*

* Received by the editors June 15, 1981, and in final revised form January 26, 1982. This research was supported in part by the National Science Foundation.

† Department of Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel.
‡ Department of Mathematics, University of Illinois, Urbana, Illinois 61801.
§ Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot, Israel.

Proof. $(a-bz)^{-\alpha}=a^{1-\alpha}\Sigma[(\alpha)_r/r!](b^r/a^{r+1})z^r$ and $(c-dz)^{-\alpha}=c^{1-\alpha}\Sigma[(\alpha)_r/r!](d^r/c^{r+1})z^r$ have positive coefficients. Thus for every r, both $a^{1-\alpha}b^r/a^{r+1}$ and $c^{1-\alpha}d^r/c^{r+1}$ do and, hence, does $(ac)^{1-\alpha}b^rd^r/a^{r+1}c^{r+1}$ and finally does

$$(ac)^{1-\alpha}\sum\frac{(\alpha)_r}{r!}\frac{b^rd^r}{a^{r+1}c^{r+1}}=(a(x,y)c(z,u)-b(x,y)d(z,u))^{-\alpha}.$$

Take $a(x,y)=1-x-y$, $b(x,y)=x+y-4xy$, $c(z,u)=1-z-u$d, $d(z,u)=z+u-4zu$. The hypotheses of Proposition 2 are satisfied (for $\alpha\geq 1$) by virtue of the above discussion, (with $\lambda=\frac{1}{2}$, $x\leftarrow 2x$, $y\leftarrow 2y$). Thus, we have an elementary proof of Koornwinder's [5] result that $[1-x-y-z-u+4(xyu+xyz+xzu+yzu)-16xyzu]^{-\alpha}$ has positive coefficients for $\alpha\geq 1$.

3. A short proof of a result of Askey and Gasper. It follows from the above that $(1-x-y-z+4xyz)^{-\beta}$ has positive coefficients for $\beta\geq 1$. Askey and Gasper [2] extended this result to $\beta\geq(\sqrt{17}-3)/2$. This can be obtained quite simply by an extension of a method used in [3].

Suppose that $\beta>(\sqrt{17}-3)/2$. Write $R=1-x-y-z+4xyz$, it is readily seen that

$$\frac{\partial}{\partial x}R^{-\beta}=(1+2z)\left[x\frac{\partial}{\partial x}-y\frac{\partial}{\partial y}+z\frac{\partial}{\partial z}+\beta\right]R^{-\beta}+2\left(y\frac{\partial}{\partial y}-z\frac{\partial}{\partial z}\right)R^{-\beta}.$$

Substitute $R^{-\beta}=\Sigma D_{a+1,b,c}x^{a+1}y^bz^c$ above, compare coefficients of $x^ay^bz^c$, and set $a\leftarrow a-1$ to get

$$aD_{a,b,c}=(a+b-c+\beta-1)D_{a-1,b,c}+2(a-b+c-2+\beta)D_{a-1,b,c-1}.$$

Now, by symmetry, it is enough to prove positivity for $a\geq b\geq c$. The coefficients of the above recurrence are positive if $a\geq b\geq c>1$ and the result will follow by induction if $D_{a,a,1}\geq 0$ for all a. Now

$$D_{a,a,1}=\frac{\beta(\beta+1)\cdots(\beta+2a-2)}{(a-1)!^2}\left[\frac{(\beta+2a-1)(\beta+2a)}{a^2}-4\right].$$

But $(\beta+2a-1)(\beta+2a)-4a^2=\beta^2-\beta+2a(2\beta-1)$ increases with a since $\beta\cong 0.56>0.5$ and $D_{1,1,1}=\beta(\beta^2+3\beta-2)>0$, so the result follows.

4. Does $(1-(x_1+\cdots+x_n)+n!x_1\cdots x_n)^{-1}$ have positive power series coefficients? We have already mentioned Askey and Gasper's result that $[1-(x+y+z)+4xyz]^{-1}$ has positive power series coefficients. We are interested in A_n, the largest A for which $(1-(x_1+\cdots+x_n)+Ax_1\cdots x_n)^{-1}$ has nonnegative coefficients. Since the coefficients of $x_1\cdots x_n$ in the above expansion is $n!-A_n$, we must certainly have $A_n\leq n!$. We conjecture that for $n\geq 4$, $A_n=n!$. It may be seen that $A_n\geq(n-1)!$, i.e., that $[1-(x_1+\cdots+x_n)+(n-1)!x_1\cdots x_n]^{-1}$ has positive coefficients. The reason is that the coefficient of $x_1^{\alpha_1}\cdots x_n^{\alpha_n}$ in the above expansion has combinatorial significance, namely, it is the number of words with α_1 1's,$\cdots,\alpha_n$ n's such that no substring of n letters which ends with the letter "n" can be a permutation (e.g., with $n=4$, the six words 1234, 1324, 2134, 2314, 3124, 3214 are not allowed as subwords) (see Zeilberger [7] for details).

Let us state:

PROPOSITION 3. *Let* $(1-(x_1+\cdots+x_n)+n!x_1\cdots x_n)^{-1}=\Sigma A_{\alpha_1,\cdots,\alpha_n}x_1^{\alpha_1}\cdots x_n^{\alpha_n}$. *If* $A_{r,\cdots,r}\geq 0$ *for all* r, *then* $A_{\alpha_1\cdots\alpha_n}\geq 0$ *for all* $(\alpha_1,\cdots,\alpha_n)\in N^n$.

The proof is rather long and we omit it here. Note that

$$A^{(n)}_{r,\dots,r} = \sum_{j=0}^{r} (-1)^j \frac{(rn-(n-1)j)!(n!)^j}{(r-j)!^n j!},$$

and it would therefore suffice to show that this binomial sum is positive. This has been verified by computer for $n=4$ and $1 \le r \le 220$. In this range $A^{(4)}_{r,\dots,r}$ increases monotonically and appears to have exponential growth. This supports our conjecture.

Acknowledgment. Many thanks are due to Gilad Bandel for his programming.

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] R. ASKEY AND G. GASPER, *Convolution structures of Laguerre polynomials*, J. Analyse Math., 31 (1977), pp. 48–68.

[3] J. GILLIS AND J. KLEEMAN, *A combinatorial proof of a positivity result*, Math. Proc. Camb. Phil. Soc., 86 (1979), pp. 13–19.

[4] M. E. H. ISMAIL AND M. V. TAMHANKAR, *A combinatorial approach to some positivity problems*, SIAM J. Math. Anal., 10 (1979), pp. 478–485.

[5] T. KOORNWINDER, *Positivity proofs for linearization and connection coefficients of orthogonal polynomials satisfying an addition theorem*, J. London Math. Soc., 2 (1978), pp. 101–114.

[6] G. SZEGO, *Über gewisse Potenzeihen mit lauter positiven Koeffizienten*, Math. Z., 37 (1933), pp. 674–688.

[7] D. ZEILBERGER, *Enumeration of words by their number of mistakes*, Discrete Math., 34 (1981), pp. 89–91.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0015 $01.25/0

LEGENDRE TYPE POLYNOMIALS UNDER AN INDEFINITE INNER PRODUCT*

ANGELO B. MINGARELLI† AND ALLAN M. KRALL‡

Abstract. The polynomials which are orthogonal with respect to the indefinite inner product

$$\langle f,g\rangle=\int_{-1}^{1} f(x)\overline{g(x)}\frac{\alpha}{2}dx+\frac{1}{2}f(1)\overline{g(1)}+\frac{1}{2}f(-1)\overline{g(-1)}$$

with $\alpha<0$ are shown to span the Pontryagin (Krein) space generated by the inner product. The polynomials are eigenfunctions associated with a selfadjoint, fourth order differential operator.

Introduction. In a recent article [2] it was shown that the Legendre type polynomials

$$P_n^{(\alpha)}(x)=\sum_{k=0}^{[n/2]}\frac{(-1)^k(2n-2k)!\left(\alpha+\dfrac{n(n-1)}{2}+2k\right)x^{n-2k}}{2^n k!(n-k)!(n-2k)!}$$

which satisfy a fourth order differential equation of the form

$$ly=\lambda_n y,\qquad n=0,1,\cdots,$$

where

$$ly=\left((x^2-1)^2y''\right)''+4\left(\left(\alpha(x^2-1)-2\right)y'\right)'$$

and

$$\lambda_n=8\alpha n+(4\alpha+12)n(n-1)+8n(n-1)(n-2)+n(n-1)(n-2)(n-3)$$

are orthogonal with respect to the Stieltjes measure ψ, given by

$$\psi(x)=\begin{cases}-\dfrac{1+\alpha}{2}, & x\le -1,\\[2mm] \dfrac{\alpha x}{2}, & -1<x<1,\\[2mm] \dfrac{1+\alpha}{2}, & 1\le x.\end{cases}$$

The polynomials have a number of additional properties associated with orthogonal polynomials. In particular it was shown that

$$\int_{-1-1}^{1+}P_n^{(\alpha)}(x)^2d\psi=\alpha\left(\alpha+\frac{(n-1)(n)}{2}\right)\left(\alpha+\frac{(n+1)(n+2)}{2}\right)/(2n+1),$$

that the set $\{P_n^{\alpha}\}_{n=0}^{\infty}$ spans the Hilbert space H generated by ψ, and that l gives rise to a selfadjoint operator in H.

The purpose of this article is to show that when $\alpha<0$ and is not the negative of a triangular number, ψ generates an indefinite inner product space K, which is a Pontryagin (Krein) space [1] with rank of positivity 2, the polynomials $\{P_n^{(\alpha)}\}_{n=0}^{\infty}$ span K, and l gives rise to a selfadjoint operator A. For a brief discussion of these spaces we refer the reader to [4], see also [1, Chap. IX].

* Received by the editors March 16, 1981.

† Department of Mathematics, The University of Ottawa, Ottawa, Ontario K1N9B4, Canada.
‡ Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.

When α is negative triangular, there does not exist a polynomial $P_n^{(\alpha)}$ exactly of degree n for each $n=0,1,\cdots$, and the space K is degenerate. When $\alpha=0$, K is also degenerate.

The polynomials. We assume that for some $N>0$

$$-\frac{N(N+1)}{2}<\alpha<-\frac{(N-1)N}{2},$$

so that α is not the negative of a triangular number. We then note that the polynomial $P_n^{(\alpha)}(x)$ is exactly of degree n and that the formulas $lP_n^{(\alpha)}=\lambda_n P_n^{(\alpha)}$ and

$$\left\langle P_n^{(\alpha)}, P_n^{(\alpha)}\right\rangle=\alpha\left(\alpha+\frac{(n-1)(n)}{2}\right)\left(\alpha+\frac{(n+1)(n+2)}{2}\right)/(2n+1)$$

still hold [2]. Since

$$\alpha+\frac{(n+1)(n+2)}{2}>\frac{1}{2}(n+N+2)(n-N+1)>0$$

when $n\geq N-1$, and

$$\alpha+\frac{(n-1)(n)}{2}>\frac{1}{2}(n+N)(n-N-1)>0$$

when $n\geq N+1$, we have

$$\left\langle P_j^{(\alpha)}, P_j^{(\alpha)}\right\rangle\begin{cases}<0, & j=0,\cdots,N-2,\\ >0, & j=N-1,N,\\ <0, & j=N+1,\cdots.\end{cases}$$

If we let K be the indefinite inner product space generated by $\langle\cdot,\cdot\rangle$ then K admits a Hilbert majorant given by

$$[f,g]=\int_{-1}^{1} f(x)\overline{g(x)}\frac{|\alpha|}{2}dx+\frac{1}{2}f(1)\overline{g(1)}+\frac{1}{2}f(-1)g(-1).$$

By [1, p. 89], $K=K^+\oplus K^0\oplus K^-$, where K^+ is a positive definite subspace, K^0 is neutral, and K^- is negative definite.

LEMMA 1. $K^0=\{0\}$. ($K=K^+\oplus K^-$.)

Proof. If f is in K^0, then f is orthogonal to all of K. In particular, f is orthogonal to x^n, for all $n=0,1,\cdots$. An argument similar to that found in [2] shows $f\equiv 0$. Thus $K^0=\{0\}$, and $K=K^+\oplus K^-$. □

Now let

$$P^+=\operatorname{span}\left\{P_n^{(\alpha)}:\left\langle P_n^{(\alpha)}, P_n^{(\alpha)}\right\rangle>0\right\},$$
$$P^-=\operatorname{span}\left\{P_n^{(\alpha)}:\left\langle P_n^{(\alpha)}, P_n^{(\alpha)}\right\rangle<0\right\}$$

and let $P=P^+\oplus P^-$.

LEMMA 2. $K=P^+\oplus P^-$.

Proof. By [1, p. 104] P is orthocomplemented in K. If f is in $P^\perp$, then by the argument of Lemma 1, $f\equiv 0$. Hence $K=P^+\oplus P^-$. □

THEOREM 1. *K is a (nondegenerate decomposable) Pontryagin (Krein) space spanned by $\{P_n^{(\alpha)}\}_{n=0}^{\infty}$, and $K=P^+\oplus P^-$.*

COROLLARY. *If f is in K, then*

$$f=\sum_{n=0}^{\infty} c_n P_n^{(\alpha)},$$

where

$$c_n=\frac{\langle f,P_n^{(\alpha)}\rangle}{\langle P_n^{(\alpha)},P_n^{(\alpha)}\rangle}.$$

The differential operator. While the differential expression l is formally selfadjoint, the boundary value problem

$$ly=\lambda y,$$
$$8\alpha y'(1)=\lambda y(1), \qquad -8\alpha y'(-1)=\lambda y(-1),$$

which is required to show symmetry in Green's formula, involves λ-dependent boundary conditions. Consequently it is convenient to express K in a slightly different form in order to fully exhibit the full role played by ± 1.

We denote by $\mathcal{K}$ the indefinite inner product space $L^2(-1,1)\otimes\mathcal{C}\otimes\mathcal{C}$, where for $F=(f(x),f_1,f_{-1})^T$ and $G=(g(x),g_1,g_{-1})^T$ in

$$\langle F,G\rangle_{\mathcal{K}}=\int_{-1}^{1} f(x)\overline{g(x)}\frac{\alpha}{2}dx+\frac{1}{2}f_1\overline{g_1}+\frac{1}{2}f_{-1}\overline{g_{-1}}.$$

It is evident that K is isomorphic to $\mathcal{K}$. The operator A is defined as follows:

Let D_A consist of those elements $Y=(y(x),y_1,y_{-1})^T$ satisfying:

1. y is in $L^2(-1,1)$,
2. y', y'', y''' exist and y''' is absolutely continuous,
3. ly exists almost everywhere and is in $L^2(-1,1)$,
4. $y_1=y(1)$,
5. $y_{-1}=y(-1)$.

We then define A by setting

$$AY=\begin{pmatrix} ly \\ ly(1) \\ ly(-1) \end{pmatrix}=\begin{pmatrix} ly \\ 8ly'(1) \\ -8ly'(-1) \end{pmatrix}.$$

Green's formula establishes that A is symmetric. Further, an argument similar to that in [3] establishes:

THEOREM 2. *A is selfadjoint in $\mathcal{K}$.*

THEOREM 3. *The spectrum of A is real and discrete. It consists only of eigenvalues*; $\sigma_p(A)=\{\lambda_n\}_{n=0}^{\infty}$.

Proof. Let $Y_n=(P_n^{(\alpha)}(x),\alpha,(-1)^n\alpha)^T$, $n=0,1,\cdots$. Then for F in $\mathcal{K}$,

$$F=\sum_{n=0}^{\infty} C_nY_n,$$

where $C_n=\langle F,Y_n\rangle_{\mathcal{K}}/\langle Y_n,Y_n\rangle_{\mathcal{K}}$. Thus if we attempt to solve $(A-\lambda I)Y=F$, with

$$Y=\sum_{n=0}^{\infty} B_nY_n,$$

we find $B_n=C_n/(\lambda-\lambda_n)$. When $\lambda\neq\lambda_n$, $n=0,1,\cdots$, Y is in $\mathcal{K}$ and is represented by

$$Y=\sum_{n=0}^{\infty}\left[\frac{C_n}{\lambda-\lambda_n}\right]Y_n.$$

Thus when $\lambda\notin\{\lambda_n\}_{n=0}^{\infty}$, $(A-\lambda I)^{-1}$ exists, and λ is in the resolvent set. □

In a similar way we can also establish

THEOREM 4. *For Y in D_A, $AY=\sum_{n=0}^{\infty}\lambda_n C_n Y_n$, where $C_n=\langle Y, Y_n\rangle_{\mathcal{K}}/\langle Y_n, Y_n\rangle_{\mathcal{K}}$. Further, Y is in D_A if and only if $\sum_{n=0}^{\infty} n^3\lambda_n^2|C_n|^2<\infty$.*

REFERENCES

[1] J. BOGNAR, *Indefinite inner product spaces*, Springer-Verlag, New York, 1974.

[2] A. M. KRALL, *Orthogonal polynomials satisfying a fourth order differential equation*, C.R. Math. Rep. Acad. Sci. Canada, 1 (1979), pp. 219–222.

[3] _____, *Orthogonal polynomials satisfying fourth order differential equations*, Proc. Roy. Soc. Edinburgh, 87A (1981), pp. 271–288.

[4] _____, *Laguerre polynomial expansions in indefinite inner product spaces*, J. Math. Anal. Appl., 70 (1979), pp. 267–279.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0016 $01.25/0

LINEARIZATION OF THE PRODUCT OF ASSOCIATED LEGENDRE POLYNOMIALS*

RUPERT LASSER†

Abstract. The linearization coefficients for the associated Legendre polynomials are found in a sufficiently explicit form so their nonnegativity can be proven.

The linearization problem for an arbitrary orthogonal polynomial sequence $\{P_n(x)\}$ is the problem of finding the coefficients $b(m,n,n+m-k)$ in the expansion

$$P_m(x)\cdot P_n(x)=\sum_{k=0}^{2m} b(m,n,n+m-k)P_{n+m-k}(x), \tag{1}$$

where $m\le n$ (see [2, Chap. 5]). We shall present a solution for the associated Legendre polynomials which can be defined by means of their recurrence relation:

$$P_{-1}(\nu;x)=0, \qquad P_0(\nu;x)=1,$$
$$P_{n+1}(\nu;x)=xP_n(\nu;x)-\frac{(n+\nu)^2}{4(n+\nu)^2-1}P_{n-1}(\nu;x), \tag{2}$$

where we assume that $\nu>-\frac{1}{2}$ (see [5, Chap. 6, §12] or [3]). We prefer to use monic polynomials, i.e., $P_n(x)=x^n+\cdots$. Thus our notation differs from that in [3] and [5]. $P_n(0;x)$ are the Legendre polynomials, and for $\nu\in\mathbb{N}$ the polynomials $P_n(\nu;x)$ are their numerator polynomials of order ν (see [5, p. 87]). In 1878 F. Neumann [6] and J. C. Adams [1] found the linearization formula for the Legendre polynomials (compare [2, (5.5)]).

Now fix $\nu>-\frac{1}{2}$. Our method of proof is induction, where we use the following recurrence relation for $b(m,n,n+m-k)$, $m\le n$:

$$b(m,n,n+m-k)=0 \quad \text{for } k=1,3,\cdots,2m-1, \qquad b(0,n,n)=1,$$
$$b(1,n,n+1)=1, \qquad b(1,n,n-1)=\frac{(n+\nu)^2}{4(n+\nu)^2-1}$$

and for $m=2,3,\cdots$

$$b(m,n,n+m)=1, \qquad b(m,n,n-m)=b(m-1,n,n-m+1)b(1,n-m+1,n-m),$$
$$\begin{aligned} b(m,n,n+m-2k)=&\,b(m-1,n,n+m-1-2k)\\ &+b(m-1,n,n+m+1-2k)b(1,n+m+1-2k,n+m-2k)\\ &-b(m-2,n,n+m-2k)b(1,m-1,m-2)\\ &\qquad\qquad \text{for } k=1,2,\cdots,m-1. \end{aligned} \tag{3}$$

* Received by the editors September 24, 1981, and in revised form December 20, 1981.

† Institut für Mathematik der Technischen Universität München, Arcisstrasse 21, D-8000 München 2, Federal Republic of Germany.

One can easily deduce (3) from the associativity of the product $(xP_{m-1}(x))P_n(x)=x(P_{m-1}(x)P_n(x))$. First we calculate certain auxiliary constants $c(m,n,n+m-2k)$, $m\leq n$, which are defined recursively in a similar way:

$$c(1,n,n+1)=1, \qquad c(1,n,n-1)=b(1,n,n-1)$$

and for $m=2,3,\cdots$

$$c(m,n,n-m)=1, \qquad c(m,n,n-m)=b(m,n,n-m),$$

(4)

$$\begin{aligned} c(m,n,n+m-2k)&=c(m-1,n,n+m-1-2k)\\ &\quad+c(m-1,n,n+m+1-2k)\\ &\qquad\cdot c(1,n+m+1-2k,n+m-2k)\\ &\quad-c(m-2,n,n+m-2k)\,\frac{(m-1)^2}{4(m-1)^2-1} \end{aligned}$$

for $k=1,2,\cdots,m-1$.

Further denote

$$B_m=\frac{(1/2)_m 2^m}{m!}=\frac{1\cdot 3\cdot 5\cdot\cdots\cdot 2m-1}{m!},$$

$$A_m=\frac{(\nu+1/2)_m 2^m}{(\nu+1)_m}=\frac{(2\nu+1)(2\nu+3)\cdots(2\nu+2m-1)}{(\nu+1)_m},$$

$$(a)_m=a(a+1)\cdots(a+m-1), \qquad (a)_0=1.$$

PROPOSITION 1. *For $m,n\in\mathbb{N}$, $m\leq n$, $k=0,1,\cdots,m$ the following identity holds*:

$$c(m,n,n+m-2k)=\frac{B_{m-k}B_kA_{n-k}A_{n+m-2k}(2n+2m-4k+1+2\nu)}{B_mA_nA_{n+m-k}(2n+2m-2k+1+2\nu)}.$$

Proof. The assertion follows immediately for $k=0$ and $k=m$. Now we shall use induction on m. Let $2\leq m\leq n$ and $k\in\{1,2,\cdots,m-1\}$. By the induction assumption we obtain

$$\begin{aligned} c(m,n,n+m-2k)&=\frac{B_{m-k}B_kA_{n-k}A_{n+m-2k}}{B_mA_nA_{n+m-k}}\,\frac{(m-k)(2m-1)(\nu+n+m-2k)}{(2m-2k-1)m(\nu+n+m-k)}\\ &\quad+\frac{k(2m-1)(2\nu+2n-2k+1)(\nu+n+m-2k+1)}{(2k-1)m(\nu+n-k+1)(2\nu+2n+2m-2k+1)}\\ &\quad-\frac{(m-k)k(m-1)(2\nu+2n-2k+1)(2\nu+2n+2m-4k+1)}{(2m-2k-1)(2k-1)m(\nu+n-k+1)(\nu+n+m-k)}. \end{aligned}$$

Now a direct computation shows that the first term minus the third term in the above brackets is equal to $-(m-k)(\nu+n-2k+1)/(m(2k-1)(\nu+n-k+1))$. Thus

$$
\begin{aligned}
&c(m,n,n+m-2k)\\
&=\frac{B_{m-k}B_kA_{n-k}A_{n+m-2k}}{B_mA_nA_{n+m-k}}\\
&\cdot\left(\frac{k(2m-1)(2\nu+2n-2k+1)(\nu+n+m-2k+1)}{(2k-1)m(\nu+n-k+1)(2\nu+2n+2m-2k+1)}\right.\\
&\qquad\left.-\frac{(m-k)(\nu+n-2k+1)(2\nu+2n+2m-2k+1)}{(2k-1)m(\nu+n-k+1)(2\nu+2n+2m-2k+1)}\right)\\
&=\frac{B_{m-k}B_kA_{n-k}A_{n+m-2k}}{B_mA_nA_{n+m-k}}\cdot\frac{2(\nu+n-k)^2+(2m-2k+3)(\nu+n-k)+(2m-2k+1)}{(\nu+n-k+1)(2\nu+2n+2m-2k+1)}\\
&=\frac{B_{m-k}B_kA_{n-k}A_{n+m-2k}}{B_mA_nA_{n+m-k}}\frac{(2\nu+2n+2m-4k+1)}{(2\nu+2n+2m-2k+1)}.
\end{aligned}
$$

□

Define the constants $R_{m,k}$, $m\in\mathbb{N}\cup\{0\}$, $k=0,1,\cdots,m$ by

$$
\begin{aligned}
&(5)\qquad R_{m,0}=1 \quad \text{if } m\in\mathbb{N}\cup\{0\}, \qquad R_{m,m}=0 \quad \text{if } m\in\mathbb{N},\\
&R_{m,k}=R_{m-1,k}+R_{m-2,k-1}\left(\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}-b(1,m-1,m-2)\right)\\
&\qquad+\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}(R_{m-1,k-1}-R_{m-2,k-1}) \quad \text{if } k=1,\cdots,m-1.
\end{aligned}
$$

THEOREM. *Let $m,n\in\mathbb{N}$, $m\le n$. Then*

$$
\begin{aligned}
&P_m(\nu;x)\cdot P_n(\nu;x)=\sum_{k=0}^{m}b(m,n,n+m-2k)P_{n+m-2k}(\nu;x),\\
&b(m,n,n+m-2k)=\sum_{j=0}^{\min(k-1,m-k)}R_{m,j}c(m-2j,n,n+m-2k)+R_{m,k},
\end{aligned}
$$

where $R_{m,j}$ is defined in (5) *above and $c(l,n,n+l-2k)$ is calculated in Proposition* 1.

Proof. Obviously the assertion holds for $k=0$ and $k=m$. Define in addition to (4) $c(m,n,n+m-2k)=0$ for $k\in\mathbb{N}$, $k>m$. Then the three-fold recurrence formula of (4) holds for each $k\in\mathbb{N}$. Thus it is sufficient to prove

$$
b(m,n,n+m-2k)=\sum_{j=0}^{k-1}R_{m,j}c(m-2j,n,n+m-2k)+R_{m,k}, \qquad k=1,\cdots,m-1.
$$

We use induction on m. By the induction assumption and (3) one obtains

$$
\begin{aligned}
&b(m,n,n+m-2k)\\
&\quad=\sum_{j=0}^{k-1} R_{m-1,j}c(m-1-2j,n,n+m-1-2k)+R_{m-1,k}\\
&\qquad+\sum_{j=0}^{k-2} R_{m-1,j}c(m-1-2j,n,n+m-1-2(k-1))\\
&\qquad\qquad\cdot c(1,n+m-1-2(k-1),n+m-2k)\\
&\qquad+R_{m-1,k-1}b(1,n+m-1-2(k-1),n+m-2k)\\
&\qquad-\sum_{j=0}^{k-2} R_{m-2,j}c(m-2-2j,n,n+m-2-2(k-1))c(1,m-1,m-2)\\
&\qquad-R_{m-2,k-1}c(1,m-1,m-2)\\
&\quad=S_1+S_2,
\end{aligned}
$$

where

$$
\begin{aligned}
S_1=\sum_{j=0}^{k-2} R_{m-1,j}\Bigg[&c(m-1-2j,n,n+m-1-2k)\\
&+c(m-1-2j,n,n+m-1-2(k-1))\\
&\quad\cdot c(1,n+m-1-2(k-1),n+m-2k)\\
&-c(m-2-2j,n,n+m-2-2(k-1))\,\frac{(m-2j-1)^2}{4(m-2j-1)^2-1}\Bigg]\\
+\sum_{j=0}^{k-2}&(R_{m-1,j}-R_{m-2,j})c(m-2-2j,n,n+m-2-2(k-1))\,\frac{(m-2j-1)^2}{4(m-2j-1)^2-1}\\
+\sum_{j=0}^{k-2}&R_{m-2,j}c(m-2-2j,n,n+m-2-2(k-1))\\
&\cdot\left(\frac{(m-2j-1)^2}{4(m-2j-1)^2-1}-b(1,m-1,m-2)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
S_2=&R_{m-1,k-1}(c(m-1-2(k-1),n,n+m-1-2k)\\
&+c(1,n+m-1-2(k-1),n+m-2k))+R_{m-1,k}-R_{m-2,k-1}\cdot\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}\\
&+R_{m-2,k-1}\left(\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}-b(1,m-1,m-2)\right).
\end{aligned}
$$

Now by (4) and $c(m-1-2(k-1),n,n+m-1-2(k-1))=1$ we see that

$$\begin{aligned}S_2 &= R_{m-1,k-1}c(m-2(k-1),n,n+m-2k)\\ &\quad+\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}(R_{m-1,k-1}-R_{m-2,k-1})+R_{m-1,k}\\ &\quad+R_{m-2,k-1}\left(\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}-b(1,m-1,m-2)\right)\\ &= R_{m,k}+R_{m-1,k-1}c(m-2(k-1),n,n+m-2k).\end{aligned}$$

Further, by (4)

$$\begin{aligned}S_1 &= \sum_{j=0}^{k-2} R_{m-1,j}c(m-2j,n,n+m-2k)\\ &\quad+\sum_{j=1}^{k-1}(R_{m-1,j-1}-R_{m-2,j-1})c(m-2j,n,n+m-2k)\frac{(m-2j+1)^2}{4(m-2j+1)^2-1}\\ &\quad+\sum_{j=1}^{k-1} R_{m-2,j-1}c(m-2j,n,n+m-2k)\left(\frac{(m-2j+1)^2}{4(m-2j+1)^2-1}-b(1,m-1,m-2)\right).\end{aligned}$$

Thus $b(m,n,n+m-2k)=S_1+S_2=\sum_{j=0}^{k-1} R_{m,j}c(m-2j,n,n+m-2k)+R_{m,k}$, and the theorem is proved completely. □

We shall show that $R_{m,k}$ is often equal to zero.

PROPOSITION 2. *Let* $m\in\mathbb{N}$. *Then* $R_{m,k}=0$ *for* $k=[m/2]+1,\cdots,m$, *where*

$$\left[\frac{m}{2}\right]=\begin{cases} l & \text{if } m=2l,\\ l & \text{if } m=2l+1.\end{cases}$$

Proof. Assume that the assertion is valid for all positive integers not greater than $m-1$. If $m=2l$, then $R_{m,k}=0$ for $k\in\{l+1,\cdots,m\}$ by (5) and the induction assumption. If $m=2l+1$, then $R_{m,k}=0$ for $k\in\{l+1,\cdots,m\}$ again, where we use in addition that $m-2k+1=0$ for $k=l+1$. □

Our formula is simple enough to deduce that the coefficients $b(m,n,n+m-2k)$ are nonnegative if $\nu\geq 0$. This fact is very important (compare [2, Ch. 5]).

COROLLARY. *If* $\nu\geq 0$ *and*

$$P_m(\nu;x)\cdot P_n(\nu;x)=\sum_{k=0}^{m} b(m,n,n+m-2k)P_{n+m-2kd}(\nu;x),$$

then $b(m,n,n+m-2k)\geq 0$.

Proof. We have to prove that the constants $R_{m,k}$, $k=1,\cdots,[m/2]$ are nonnegative. We shall again use induction on m. Let $m\geq 2$. Since $\nu\geq 0$ we have

$$b(1,m-1,m-2)\leq\frac{(m-1)^2}{4(m-1)^2-1}\leq\frac{(m-2k+1)^2}{4(m-2k+1)^2-1}\quad\text{for } k=1,\cdots,\left[\frac{m}{2}\right].$$

Thus it is sufficient to prove that $R_{m-1,k-1}-R_{m-2,k-1}\geq 0$. Using (5) again this follows if $R_{m-2,k-2}-R_{m-3,k-2}\geq 0$. Continuing in this fashion we see that $R_{m-1,k-1}-R_{m-2,k-1}$ is the sum of nonnegative numbers. This completes the proof of the corollary. □

Remark. Denote $d_n=n^2/(4n^2-1)-b(1,n,n-1)$. Then $R_{m,1}=\sum_{k=1}^{m-1} d_k$.

Finally we point out that our nonnegativity results do not follow from the general theorem of Askey [2, Thm. 5.2] for any value of ν. This differs from the situation in [4], where Askey's theorem was applied to many associated continuous q-ultraspherical polynomials.

REFERENCES

[1] J. C. ADAMS, *On the expression for the product of any two Legendre's coefficients by means of a series of Legendre's coefficients*, Proc. Roy. Soc. London, 27 (1878), pp. 63–71.

[2] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] P. BARRUCAND AND D. DICKINSON, *On the associated Legendre polynomials*, in Proc. Conference on Orthogonal Expansions and their Continuous Analogues, D. Haimo, ed., Southern Illinois Univ. Press, Edwardsville, IL, 1968, pp. 43–50.

[4] J. BUSTOZ AND M. E. H. ISMAIL, *The associated ultraspherical polynomials and their q-analogues*, to appear.

[5] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[6] F. E. NEUMANN, *Beiträge zur Theorie der Kugelfunktionen*, Leipzig, 1878.

SIAM J. MATH. ANAL.
Vol. 14, No. 2, March 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1402-0017 $01.25/0

POSITIVITY OF THE POISSON KERNEL FOR THE CONTINUOUS q-ULTRASPHERICAL POLYNOMIALS*

GEORGE GASPER† AND MIZAN RAHMAN‡

Abstract. Rogers' [Proc. London Math. Soc., 24 (1893), pp. 117–179] bilinear generating function for the continuous q-Hermite polynomials

$$\frac{(t^2;q)_\infty}{\left|(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty\right|^2}=\sum_{n=0}^{\infty}\frac{H_n(\cos\theta|q)H_n(\cos\psi|q)}{(q;q)_n}t^n,$$

where $(q;q)_n=(1-q)(1-q^2)\cdots(1-q^n)$ and $(a;q)_\infty=(1-a)(1-aq)(1-aq^2)\cdots$, is extended to the continuous q-ultraspherical polynomials $C_n(x;\beta|q)$ and used to give conditions for the positivity of the Poisson kernel for these polynomials. Related bilinear generating functions are also considered.

1. Introduction. The continuous q-ultraspherical polynomials $C_n(x;\beta|q)$, which can be defined by the generating function

$$\frac{(\beta te^{i\theta};q)_\infty(\beta te^{-i\theta};q)_\infty}{(te^{i\theta};q)_\infty(te^{-i\theta};q)_\infty}=\sum_{n=0}^{\infty}C_n(x;\beta|q)t^n,\qquad |t|<1,\quad |q|<1,\tag{1.1}$$

where $x=\cos\theta$ and $(a;q)_\infty=(1-a)(1-aq)(1-aq^2)\cdots$, were introduced by L. J. Rogers in his work [12]–[14] on the Rogers–Ramanujan identities, and have recently been studied by Askey and Ismail [2] and Bressoud [5]. Rogers [11] showed that the continuous q-Hermite polynomials (in the notation of [2])

$$H_n(x|q)=(q;q)_nC_n(x;0|q)$$

where $(a;q)_n=(1-a)(1-aq)\cdots(1-aq^{n-1})$, have a bilinear generating function of the form

$$\frac{(t^2;q)_\infty}{\left|(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty\right|^2}=\sum_{n=0}^{\infty}\frac{H_n(\cos\theta|q)H_n(\cos\psi|q)t^n}{(q;q)_n},\tag{1.2}$$

which is a q-analogue of Mehler's formula for the Hermite polynomials. For Mehler's formula see Szegö [16, Problem 23] and Watson [17], and for some recent proofs of (1.2) see Bressoud [4] and Carlitz [6], [7]. Our main aim in this paper is to derive an extension of (1.2) to the continuous q-ultraspherical polynomials and then use it to prove that the Poisson kernel (2.1) for these polynomials is nonnegative for $0\le\beta<1$, $-1<q<1$, $-1<t<1$, $-1\le x,y\le 1$ and for some other cases. In addition, a related generating function is considered, and a new transformation formula for a certain ${}_4\phi_3$ basic hypergeometric function is derived.

*Received by the editors December 18, 1981, and in revised form March 29, 1982.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60201. The research of this author was supported in part by the National Science Foundation under grant MCS-8002507.

‡Department of Mathematics, Carleton University, Ottawa, Canada K1S5B6. The research of this author was supported in part by the Natural Sciences and Engineering Research Council (Canada) under grant A6197.

2. Extensions of (1.2). First observe that from the orthogonality relation [2,(6.6)]

$$\int_{-1}^{1} H_n(x|q)H_m(x|q)|(e^{2i\theta};q)_\infty|^2\frac{dx}{(1-x^2)^{1/2}}=\frac{2\pi(q;q)_n}{(q;q)_\infty}\delta_{m,n},$$

it follows that the right-hand side of (1.2) is a positive multiple of the Poisson kernel for $H_n(x|q)$ (and thus this kernel is clearly positive for $-1<t,q<1$). Now observe that since [2,(4.3)]

$$\int_{-1}^{1} C_n(x;\beta|q)C_m(x;\beta|q)\left|\frac{(e^{2i\theta};q)_\infty}{(\beta e^{2i\theta};q)_\infty}\right|^2\frac{dx}{(1-x^2)^{1/2}}=\frac{\delta_{m,n}}{h_n(\beta|q)}$$

for $-1<\beta<1$ with

$$h_n(\beta|q)=\frac{(\beta^2;q)_\infty(q;q)_\infty}{2\pi(\beta;q)_\infty(\beta q;q)_\infty}\frac{(q;q)_n(1-\beta q^n)}{(\beta^2;q)_n(1-\beta)},$$

the Poisson kernel

$$P_t(x,y;\beta|q)=\sum_{n=0}^{\infty}h_n(\beta|q)C_n(x;\beta|q)C_n(y;\beta|q)t^n \tag{2.1}$$

is a positive constant multiple of the sum

$$K_t(x,y;\beta|q) \tag{2.2}$$

$$=\sum_{n=0}^{\infty}\frac{(q;q)_n(1-\beta q^n)}{(\beta^2;q)_n(1-\beta)}C_n(x;\beta|q)C_n(y;\beta|q)t^n,$$

and so in looking for an extension of (1.2) we are led to look for a formula for (2.2) which reduces to (1.2) when $\beta=0$.

In analogy with the Watson type formula techniques used in Bailey [3], Gasper [8] and Rahman [9], one would expect to be able to use a special case of the Watson type formula in Rahman [10] to derive the desired extension of (1.2). Unfortunately, this approach led to computational difficulties which forced us to look for another technique. The technique employed here is essentially a modification of that used by Carlitz in [6, p. 366] to prove (1.2).

Setting $x=\cos\theta$, $y=\cos\psi$ and using [2,(3.1)]

$$C_n(y;\beta|q)=\sum_{m=0}^{n}\frac{(\beta;q)_m(\beta;q)_{n-m}}{(q;q)_m(q;q)_{n-m}}e^{i(n-2m)\psi}$$

and the inversion of Rogers' linearization formula [2,(4.19)]

$$C_{m+n}(x;\beta|q)=\sum_{j=0}^{\min(m,n)}a(j,m,n)C_{m-j}(x;\beta|q)C_{n-j}(x;\beta|q),$$

where

$$a(j,m,n)=\frac{(\beta;q)_{m+n}(q;q)_m(q;q)_n}{(q;q)_{m+n}(\beta;q)_m(\beta;q)_n}$$

$$\cdot\frac{(q^{-m-n}\beta^{-2};q)_j(1-q^{2j-m-n}\beta^{-2})(\beta^{-1};q)_j}{(q;q)_j(1-q^{-m-n}\beta^{-2})(q^{1-m-n}\beta^{-1};q)_j}\left(\frac{\beta^2}{q}\right)^j,$$

we find that

$$K_t(x,y;\beta|q)$$

$$=\sum_{n=0}^{\infty}\frac{(q;q)_n(1-\beta q^n)}{(\beta^2;q)_n(1-\beta)}t^nC_n(x;\beta|q)\sum_{m=0}^{n}\frac{(\beta;q)_m(\beta;q)_{n-m}}{(q;q)_m(q;q)_{n-m}}e^{i(n-2m)\psi}$$

$$=\sum_{m,n\geq 0}t^{m+n}\frac{(q;q)_{m+n}(1-\beta q^{m+n})(\beta;q)_m(\beta;q)_n}{(\beta^2;q)_{m+n}(1-\beta)(q;q)_m(q;q)_n}e^{i(n-m)\psi}C_{m+n}(x;\beta|q)$$

$$=\sum_{m,n\geq j\geq 0}t^{m+n}e^{i(n-m)\psi}C_{m-j}(x;\beta|q)C_{n-j}(x;\beta|q)$$

$$\cdot\frac{(\beta;q)_{m+n}(q^{-m-n}\beta^{-2};q)_j(\beta^{-1};q)_j(1-q^{2j-m-n}\beta^{-2})(1-\beta q^{m+n})}{(\beta^2;q)_{m+n}(q^{1-m-n}\beta^{-1};q)_j(q;q)_j(1-q^{-m-n}\beta^{-2})(1-\beta)}\left(\frac{\beta^2}{q}\right)^j$$

$$=\sum_{m,n,j\geq 0}t^{m+n+2j}e^{i(n-m)\psi}C_m(x;\beta|q)C_n(x;\beta|q)$$

$$\cdot\frac{(\beta;q)_{m+n+2j}(q^{-m-n-2j}\beta^{-2};q)_j(\beta^{-1};q)_j(1-q^{-m-n}\beta^{-2})(1-\beta q^{m+n+2j})}{(\beta^2;q)_{m+n+2j}(q^{1-m-n-2j}\beta^{-1};q)_j(q;q)_j(1-q^{-m-n-2j}\beta^{-2})(1-\beta)}\left(\frac{\beta^2}{q}\right)^j$$

$$=\sum_{m,n\geq 0}t^{m+n}e^{i(n-m)\psi}\frac{(\beta;q)_{m+n}(1-\beta q^{m+n})}{(\beta^2;q)_{m+n}(1-\beta)}C_m(x;\beta|q)C_n(x;\beta|q)$$

$$\cdot{}_4\phi_3\left[\begin{matrix}\beta q^{m+n},q^{1+(m+n)/2}\sqrt{\beta},-q^{1+(m+n)/2}\sqrt{\beta},\beta^{-1}\\ q^{(m+n)/2}\sqrt{\beta},-q^{(m+n)/2}\sqrt{\beta},\beta^2q^{1+m+n}\end{matrix};q,\beta t^2\right],$$

where a ${}_{r+1}\phi_r$ basic hypergeometric series is defined by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,z\right]=\sum_{k=0}^{\infty}\frac{(a_1;q)_k(a_2;q)_k\cdots(a_{r+1};q)_k}{(q;q)_k(b_1;q)_k\cdots(b_r;q)_k}z^k$$

whenever the series converges. Now use the transformation formula

$$(2.3)\qquad {}_4\phi_3\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b^{-1}\\ \sqrt{a},-\sqrt{a},abq\end{matrix};q,tb\right]=\frac{(t;q)_\infty(aq;q)_\infty}{(tb;q)_\infty(abq;q)_\infty}{}_2\phi_1\left[\begin{matrix}b,tb\\ tq\end{matrix};q,aq\right],$$

which is proved in §3, and the decomposition

$$\frac{(\beta^2q;q)_{m+n}}{(\beta^2;q)_{m+n}}=\frac{1}{1-\beta^2}-\frac{\beta^2}{1-\beta^2}q^{m+n},$$

to obtain

$$\begin{aligned}
&K_t(x,y;\beta|q)\\
&=\frac{(t^2;q)_\infty(\beta q;q)_\infty}{(\beta t^2;q)_\infty(\beta^2q;q)_\infty}\sum_{m,n\geq 0}t^{m+n}e^{i(n-m)\psi}C_m(x;\beta|q)C_n(x;\beta|q)\\
&\qquad\cdot\frac{(\beta^2q;q)_{m+n}}{(\beta^2;q)_{m+n}}{}_2\phi_1\left[\begin{matrix}\beta, & \beta t^2\\ & qt^2\end{matrix};q,\beta q^{1+m+n}\right]\\
&=\frac{(t^2;q)_\infty(\beta q;q)_\infty}{(\beta t^2;q)_\infty(\beta^2;q)_\infty}\sum_{j=0}^{\infty}\frac{(\beta;q)_j(\beta t^2;q)_j(\beta q)^j}{(q;q)_j(qt^2;q)_j}\\
&\qquad\cdot\left\{\left[\sum_{m=0}^{\infty}(tq^je^{-i\psi})^mC_m(x;\beta|q)\right]\left[\sum_{n=0}^{\infty}(tq^je^{i\psi})^nC_n(x;\beta|q)\right]\right.\\
&\qquad\left.-\beta^2\left[\sum_{m=0}^{\infty}(tq^{j+1}e^{-i\psi})^mC_m(x;\beta|q)\right]\left[\sum_{n=0}^{\infty}(tq^{j+1}e^{i\psi})^nC_n(x;\beta|q)\right]\right\}.
\end{aligned}$$

Hence, by (1.1), we have the following extension of (1.2)

$$\begin{aligned}
K_t(x,y;\beta|q)=&\frac{(t^2;q)_\infty(\beta q;q)_\infty}{(\beta t^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta te^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2\\
&\cdot\sum_{j=0}^{\infty}\frac{(\beta;q)_j(\beta t^2;q)_j(\beta q)^j}{(q;q)_j(qt^2;q)_j}\left\{\left|\frac{(te^{i(\theta+\psi)};q)_j(te^{i(\theta-\psi)};q)_j}{(\beta te^{i(\theta+\psi)};q)_j(\beta te^{i(\theta-\psi)};q)_j}\right|^2\right.\\
&\left.-\beta^2\left|\frac{(te^{i(\theta+\psi)};q)_{j+1}(te^{i(\theta-\psi)};q)_{j+1}}{(\beta te^{i(\theta+\psi)};q)_{j+1}(\beta te^{i(\theta-\psi)};q)_{j+1}}\right|^2\right\}.
\end{aligned}\tag{2.4}$$

Since

$$\begin{aligned}
&\left|(1-\beta tq^je^{i(\theta+\psi)})(1-\beta tq^je^{i(\theta-\psi)})\right|^2-\beta^2\left|(1-tq^je^{i(\theta+\psi)})(1-tq^je^{i(\theta-\psi)})\right|^2\\
&=(1-\beta)(1-\beta t^2q^{2j})\left[1+\beta+\beta t^2q^{2j}+\beta^2t^2q^{2j}-2\beta tq^j\{\cos(\theta+\psi)+\cos(\theta-\psi)\}\right]\\
&=(1-\beta)(1-\beta t^2q^{2j})\left[\left|1-\beta tq^je^{i(\theta-\psi)}\right|^2+\beta\left|1-tq^je^{i(\theta+\psi)}\right|^2\right]
\end{aligned}$$

and

$$1-\beta t^2q^{2j}=(1-\beta t^2)\frac{\left(q\sqrt{\beta t^2};q\right)_j\left(-q\sqrt{\beta t^2};q\right)_j}{\left(\sqrt{\beta t^2};q\right)_j\left(-\sqrt{\beta t^2};q\right)_j},$$

we can also write (2.4) in the following equivalent forms

$$K_t(x,y;\beta|q)$$

$$=\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta qt^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta te^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$\cdot\left\{\sum_{j=0}^{\infty}\frac{(\beta t^2;q)_j\left(q\sqrt{\beta t^2};q\right)_j\left(-q\sqrt{\beta t^2};q\right)_j(\beta;q)_j}{(q;q)_j\left(\sqrt{\beta t^2};q\right)_j\left(-\sqrt{\beta t^2};q\right)_j(qt^2;q)_j}\right. \tag{2.5}$$

$$\cdot\left|\frac{(te^{i(\theta+\psi)};q)_j(te^{i(\theta-\psi)};q)_j}{(\beta te^{i(\theta+\psi)};q)_{j+1}(\beta te^{i(\theta-\psi)};q)_{j+1}}\right|^2$$

$$\left.\cdot\left[|1-\beta tq^je^{i(\theta-\psi)}|^2+\beta|1-tq^je^{i(\theta+\psi)}|^2\right](\beta q)^j\right\}.$$

$$=\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta qt^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta qte^{i(\theta+\psi)};q)_\infty(\beta te^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}\beta t^2, q\sqrt{\beta t^2}, -q\sqrt{\beta t^2}, \beta, te^{i(\theta+\psi)}, te^{-i(\theta+\psi)}, te^{i(\theta-\psi)}, te^{i(\psi-\theta)}\\ \sqrt{\beta t^2}, -\sqrt{\beta t^2}, qt^2, \beta qte^{-i(\theta+\psi)}, \beta qte^{i(\theta+\psi)}, \beta te^{i(\psi-\theta)}, \beta te^{i(\theta-\psi)}\end{matrix};q,\beta q\right]$$

$$+\beta\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta qt^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta qte^{i(\theta+\psi)};q)_\infty(\beta qte^{i(\theta-\psi)};q)_\infty}{(qte^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2 \tag{2.6}$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}\beta t^2, q\sqrt{\beta t^2}, -q\sqrt{\beta t^2}, \beta, qte^{i(\theta+\psi)}, qte^{-i(\theta+\psi)}, & te^{i(\theta-\psi)}, te^{i(\psi-\theta)}\\ \sqrt{\beta t^2}, -\sqrt{\beta t^2}, qt^2, \beta qte^{-i(\theta+\psi)}, \beta qte^{i(\theta+\psi)}, & \beta qte^{i(\psi-\theta)}, \beta qte^{i(\theta-\psi)}\end{matrix};q,\beta q\right].$$

A simpler formula than (2.6) can be derived by using the Heine transformation [1, II1]

$${}_2\phi_1\left[\begin{matrix}a,b\\c\end{matrix};q,z\right]=\frac{(az;q)_\infty(b;q)_\infty}{(z;q)_\infty(c;q)_\infty}{}_2\phi_1\left[\begin{matrix}c/b,z\\az\end{matrix};q,b\right] \tag{2.7}$$

and (2.3) to obtain

$${}_4\phi_3\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b^{-1}\\ \sqrt{a},-\sqrt{a},abq\end{matrix};q,tb\right] \tag{2.8}$$

$$=\frac{(t;q)_\infty(aq;q)_\infty}{(tq;q)_\infty(abq;q)_\infty}{}_2\phi_1\left[\begin{matrix}b,tb\\tq\end{matrix};q,aq\right]$$

$$=\frac{(t;q)_\infty}{(tq;q)_\infty}{}_2\phi_1\left[\begin{matrix}q/b,aq\\abq\end{matrix};q,tb\right]$$

$$=\frac{(t;q)_\infty(aq;q)_\infty}{(tb;q)_\infty(ab;q)_\infty}{}_4\phi_3\left[\begin{matrix}tb/q,q(tb/q)^{1/2},-q(tb/q)^{1/2},b/q\\(tb/q)^{1/2},-(tb/q)^{1/2},tq\end{matrix};q,aq\right],$$

which, when used in place of (2.3), gives

$$K_t(x,y;\beta|q)=\frac{(t^2;q)_\infty(\beta q;q)_\infty}{(\beta t^2;q)_\infty(\beta^2;q)_\infty}\sum_{j=0}^{\infty}\frac{(\beta t^2/q;q)_j\left(q(\beta t^2/q)^{1/2};q\right)_j}{(q;q)_j\left((\beta t^2/q)^{1/2};q\right)_j}$$

$$\cdot\frac{\left(-q(\beta t^2/q)^{1/2};q\right)_j(\beta/q;q)_j}{\left(-(\beta t^2/q)^{1/2};q\right)_j(qt^2;q)_j}(\beta q)^j\left[\sum_{m=0}^{\infty}(tq^je^{-i\psi})^mC_m(x;\beta|q)\right]$$

(2.9)

$$\cdot\left[\sum_{n=0}^{\infty}(tq^je^{i\psi})^nC_n(x;\beta|q)\right]$$

$$=\frac{(t^2;q)_\infty(\beta q;q)_\infty}{(\beta t^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta te^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$\cdot{}_8\phi_7\left[\begin{matrix}\beta t^2/q,q(\beta t^2/q)^{1/2},-q(\beta t^2/q)^{1/2},\beta/q,te^{i(\theta+\psi)},\\ (\beta t^2/q)^{1/2},-(\beta t^2/q)^{1/2},qt^2,\beta te^{-i(\theta+\psi)},\end{matrix}\right.$$

$$\left.\begin{matrix}te^{-i(\theta+\psi)},te^{i(\theta-\psi)},te^{i(\psi-\theta)}\\ \beta te^{i(\theta+\psi)},\beta te^{i(\psi-\theta)},\beta te^{i(\theta-\psi)}\end{matrix};q,\beta q\right].$$

Moreover, this ${}_8\phi_7$ series is very well-poised and hence can be transformed via the formula

$${}_8\phi_7\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b,c,d,e,f\\ \sqrt{a},-\sqrt{a},aq/b,aq/c,aq/d,aq/e,aq/f\end{matrix};q,\frac{a^2q^2}{bcdef}\right]$$

$$=\frac{(aq;q)_\infty(aq/ef;q)_\infty(a^2q^2/bcde;q)_\infty(a^2q^2/bcdf;q)_\infty}{(aq/e;q)_\infty(aq/f;q)_\infty(a^2q^2/bcd;q)_\infty(a^2q^2/bcdef;q)_\infty}$$

(2.10)

$$\cdot{}_8\phi_7\left[\begin{matrix}a^2q/bcd,q(a^2q/bcd)^{1/2},-q(a^2q/bcd)^{1/2},aq/cd,aq/bd,\\ (a^2q/bcd)^{1/2},-(a^2q/bcd)^{1/2},aq/b,aq/c,\end{matrix}\right.$$

$$\left.\begin{matrix}aq/bc,e,f,\\ aq/d,a^2q^2/bcde,a^2q^2/bcdf,\end{matrix};q,\frac{aq}{ef}\right],$$

which is the limit case of [15,(3.4.2.4)], to yield

(2.11)

$$K_t(x,y;\beta|q)$$

$$=\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta qt^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta qte^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$\cdot{}_8\phi_7\left[\begin{matrix}\beta t^2,q(\beta t^2)^{1/2},-q(\beta t^2)^{1/2},\beta,qte^{i(\theta+\psi)},qte^{-i(\theta+\psi)},te^{i(\theta-\psi)},te^{i(\psi-\theta)}\\ (\beta t^2)^{1/2},-(\beta t^2)^{1/2},qt^2,\beta te^{-i(\theta+\psi)},\beta te^{i(\theta+\psi)},\beta qte^{i(\psi-\theta)},\beta qte^{i(\theta-\psi)}\end{matrix};q,\beta\right].$$

Notice that from both (2.6) and (2.11) it is clear that $K_t(x,y;\beta|q)$ and hence the Poisson kernel $P_t(x,y;\beta|q)$ are positive if $0\le\beta<1$, $-1<q<1$, $-1<t<1$ and $-1\le x,y\le 1$.

The above methods can also be employed to derive other bilinear generating functions. But here we shall only point out that since, as above,

$$L_t(x,y;\beta|q)\equiv\sum_{n=0}^{\infty}\frac{(q;q)_n}{(\beta^2;q)_n}C_n(x;\beta|q)C_n(y;\beta|q)t^n$$

$$(2.12)\qquad =\sum_{m,n\ge 0}t^{m+n}e^{i(n-m)\psi}C_m(x;\beta|q)C_n(x;\beta|q)$$

$$\cdot\frac{(\beta;q)_{m+n}}{(\beta^2;q)_{m+n}}{}_2\phi_1\left[\begin{matrix}\beta^{-1},\beta q^{m+n}\\ \beta^2q^{1+m+n}\end{matrix};q,\beta t^2\right],$$

and, by (2.3),

$${}_2\phi_1\left[\begin{matrix}\beta^{-1},\beta q^{m+n}\\ \beta^2q^{1+m+n}\end{matrix};q,\beta t^2\right]$$

$$(2.13)\qquad =\frac{(t^2;q)_\infty(\beta q^{m+n};q)_\infty}{(\beta t^2;q)_\infty(\beta^2q^{m+n};q)_\infty}$$

$$\cdot{}_4\phi_3\left[\begin{matrix}\beta t^2/q,q(\beta t^2/q)^{1/2},-q(\beta t^2/q)^{1/2},\beta\\ (\beta t^2/q)^{1/2},-(\beta t^2/q)^{1/2},t^2\end{matrix};q,\beta q^{m+n}\right],$$

it follows, as in the proof of (2.9), that

$$L_t(x,y;\beta|q)=\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta t^2;q)_\infty(\beta^2;q)_\infty}\left|\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta te^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$(2.14)\qquad \cdot{}_8\phi_7\left[\begin{matrix}\beta t^2/q,q(\beta t^2/q)^{1/2},-q(\beta t^2/q)^{1/2},\beta,te^{i(\theta+\psi)},te^{-i(\theta+\psi)},\\ (\beta t^2/q)^{1/2},-(\beta t^2/q)^{1/2},t^2,\beta te^{-i(\theta+\psi)},\beta te^{i(\theta+\psi)},\\ \qquad\qquad te^{i(\theta-\psi)},te^{i(\psi-\theta)}\\ \qquad\qquad \beta te^{i(\psi-\theta)},\beta te^{i(\theta-\psi)}\end{matrix};q,\beta\right],$$

which is clearly positive if $0\le\beta<1$, $-1<q<1$, $-1<t<1$ and $-1\le x,y\le 1$. Analogous to Bailey's proof in [3] of his formula for the Poisson kernel for Jacobi polynomials [3,(2.3)], formula (2.14) can also be used to prove (2.11).

3. Proof of (2.3). Letting $f(t)$ denote the left-hand side of (2.3) and using the decomposition

$$1-aq^{2j}=1-q^j+q^j(1-aq^j),$$

we find that

$$f(t)=\frac{t(b-1)}{1-abq}\,{}_2\phi_1\left[\begin{matrix}aq,qb^{-1}\\ abq^2\end{matrix};q,tb\right]+{}_2\phi_1\left[\begin{matrix}aq,b^{-1}\\ abq\end{matrix};q,tbq\right].$$

Application of (2.7) then gives

$$f(t)=\frac{t(b-1)(tq;q)_\infty(aq;q)_\infty}{(1-abq)(tb;q)_\infty(abq^2;q)_\infty}$$

$$\cdot\,{}_2\phi_1\left[\begin{matrix}bq,tb\\ tq\end{matrix};q,aq\right]+\frac{(tq;q)_\infty(aq;q)_\infty}{(tbq;q)_\infty(abq;q)_\infty}\,{}_2\phi_1\left[\begin{matrix}b,tbq\\ tq\end{matrix};q,aq\right]$$

$$=\frac{(tq;q)_\infty(aq;q)_\infty}{(tb;q)_\infty(abq;q)_\infty}\left\{t(b-1)\,{}_2\phi_1\left[\begin{matrix}bq,tb\\ tq\end{matrix};q,aq\right]+(1-tb)\,{}_2\phi_1\left[\begin{matrix}b,tbq\\ tq\end{matrix};q,aq\right]\right\}$$

$$=\frac{(t;q)_\infty(aq;q)_\infty}{(tb;q)_\infty(abq;q)_\infty}\,{}_2\phi_1\left[\begin{matrix}b,tb\\ tq\end{matrix};q,aq\right],$$

since $t(b-1)(bq;q)_j(tb;q)_j+(1-tb)(b;q)_j(tbq;q)_j=(1-t)(b;q)_j(tb;q)_j$, which completes the proof. □

4. Additional observations. In §2 we pointed out that the positivity of the Poisson kernel $P_t(x,y;\beta|q)$ for $0\le\beta<1$, $-1<q<1$, $-1<t<1$, $-1\le x,y\le 1$ follows from (2.11). It is clear from (2.9) that $P_t(x,y;\beta|q)$ is also positive when $-1<q<\beta<0$, $-1<t<1$ and $-1\le x,y\le 1$.

If $-1<\beta<0$ and $-1<t<1$, then

$$g_t(\theta,\psi,j,\beta,q)\equiv\left|1-\beta tq^j e^{i(\theta-\psi)}\right|^2+\beta\left|1-tq^j e^{i(\theta+\psi)}\right|^2$$

$$\ge|1+\beta|^2+\beta|1+1|^2=\beta^2+6\beta+1=g_1(\pi,0,0,\beta,q)\ge 0 \tag{4.1}$$

if and only if $\beta\ge 2^{3/2}-3=0.1715728\cdots$. Hence it follows from (2.5) that

$$P_t(x,y;\beta|q)\ge 0,\qquad -\le x,y\le 1,\quad -1<t<1, \tag{4.2}$$

when $2^{3/2}-3\le\beta\le 0$ and $-1<q\le 0$. This result is the best possible in the sense that if $-1<\beta<2^{3/2}-3$ then, by (2.5) and (4.1), (4.2) fails to hold for $q=0$ and hence by continuity, for sufficiently small negative q. Unfortunately, if $\beta q<0$ then the term $(\beta q)^j$ in (2.5) alternates in sign and so it is not clear from (2.5) when $P_t(x,y;\beta|q)$ is nonnegative in this case.

Since the ultraspherical polynomials are limits of the continuous q-ultraspherical polynomials, explicitly

$$C_n^\lambda(x)=\lim_{q\to 1^-}C_n(x;q^\lambda|q),$$

it is natural to look at the corresponding limit case of formula (2.11). If we apply the transformation formula (2.10) to (2.11) to get

(4.3)

$$K_t(x,y;\beta|q)$$

$$=(1-t^2)\frac{(\beta te^{i(\theta+\psi)};q)_\infty(\beta qte^{i(\theta+\psi)};q)_\infty}{(te^{i(\theta+\psi)};q)_\infty(\beta^2qte^{i(\theta+\psi)};q)_\infty}$$

$$\cdot\left|\frac{(\beta qte^{i(\theta-\psi)};q)_\infty}{(te^{i(\theta-\psi)};q)_\infty}\right|^2$$

$$\cdot{}_8\phi_7\left[\begin{matrix}\beta^2te^{i(\theta+\psi)},q(\beta^2te^{i(\theta+\psi)})^{1/2},-q(\beta^2te^{i(\theta+\psi)})^{1/2},\beta e^{2i\theta},\beta e^{2i\psi},\beta q,\\ (\beta^2te^{i(\theta+\psi)})^{1/2},-(\beta^2te^{i(\theta+\psi)})^{1/2},\beta qte^{i(\psi-\theta)},\beta qte^{i(\theta-\psi)},\beta te^{i(\theta+\psi)},\end{matrix}\right.$$

$$\left.\begin{matrix}\beta,qte^{i(\theta+\psi)}\\ \beta qte^{i(\theta+\psi)},\beta^2\end{matrix};q,te^{-i(\theta+\psi)}\right],$$

replace β by q^λ and use the fact that

$$\lim_{q\to1^-}\frac{(zq^a;q)_\infty}{(z;q)_\infty}=(1-z)^{-a},$$

we obtain the known formula

(4.4)

$$\lim_{q\to1^-}K_t(x,y;q^\lambda|q)$$

$$=\sum_{n=0}^{\infty}\frac{(n+\lambda)n!}{\lambda(2\lambda)_n}C_n^\lambda(x)C_n^\lambda(y)t^n$$

$$=\frac{1-t^2}{(1-2t\cos(\theta-\psi)+t^2)^{\lambda+1}}\,{}_2F_1\left[\begin{matrix}\lambda+1,\lambda\\ 2\lambda\end{matrix};\frac{-4t\sin\theta\sin\psi}{1-2t\cos(\theta-\psi)+t^2}\right]$$

$$=\frac{1-t^2}{(1-2t\cos\theta\cos\psi+t^2)^{\lambda+1}}\,{}_2F_1\left[\begin{matrix}(\lambda+1)/2,(\lambda+2)/2\\ \lambda+\frac12\end{matrix};\frac{4t^2\sin^2\theta\sin^2\psi}{(1-2t\cos\theta\cos\psi+t^2)^2}\right],$$

where $(a)_n=a(a+1)\cdots(a+n-1)$, and we used the quadratic transformation

$$(4.5)\qquad {}_2F_1\left[\begin{matrix}a,b\\ 2b\end{matrix};z\right]=(1-z/2)^{-a}\,{}_2F_1\left[\begin{matrix}a/2,(a+1)/2\\ b+\frac12\end{matrix};\left(\frac{z}{2-z}\right)^2\right].$$

The right-hand side of (4.4) gives the well-known result that the Poisson kernel for ultraspherical polynomials is positive for $\lambda>-\frac12$ when $-1<t<1$ and $-1\le x,y\le1$. Since the polynomials $C_n(x;\beta|q)$ are orthogonal with respect to measure which is absolutely continuous on $(-1,1)$ and has point masses at $\pm(\beta^{1/2}+\beta^{-1/2})/2$ when $1<\beta<q^{-1/2}$, $0<q<1$, this suggests the conjecture that $P_t(x,y;\beta|q)$ should be positive

for $1<\beta<q^{-1/2}$, $0<q<1$, $-1<t<1$ when $x,y\in\{-(\beta^{1/2}+\beta^{-1/2})/2\}\cup[-1,1]\cup\{(\beta^{1/2}+\beta^{-1/2})/2\}$. Observe that by setting $e^{-i\psi}=\beta^{1/2}$, $y=(e^{i\psi}+e^{-i\psi})/2=(\beta^{1/2}+\beta^{-1/2})/2$, it follows from (2.11) and the summation formula [15, (3.3.1.4)] that

(4.6)

$$
\begin{aligned}
&K_t\big(x,(\beta^{1/2}+\beta^{-1/2})/2;\beta|q\big)\\
&\quad=\frac{(t^2;q)_\infty(\beta;q)_\infty}{(\beta qt^2;q)_\infty(\beta^2;q)_\infty}\\
&\qquad\cdot\frac{(\beta^{1/2}te^{i\theta};q)_\infty(\beta^{3/2}qte^{i\theta};q)_\infty(\beta^{3/2}te^{-i\theta};q)_\infty(\beta^{1/2}qte^{-i\theta};q)_\infty}{(\beta^{-1/2}te^{i\theta};q)_\infty(\beta^{1/2}te^{i\theta};q)_\infty(\beta^{1/2}te^{i\theta};q)_\infty(\beta^{-1/2}te^{-i\theta};q)_\infty}\\
&\qquad\cdot{}_6\phi_5\left[\begin{matrix}\beta t^2, q(\beta t^2)^{1/2}, -q(\beta t^2)^{1/2}, \beta, \beta^{-1/2}qte^{i\theta}, \beta^{-1/2}te^{-i\theta}\\ (\beta t^2)^{1/2}, -(\beta t^2)^{1/2}, qt^2, \beta^{3/2}te^{-i\theta}, \beta^{3/2}qte^{i\theta}\end{matrix};q,\beta\right]\\
&\quad=(1-t^2)\frac{(\beta^{1/2}qte^{i\theta};q)_\infty(\beta^{1/2}qte^{-i\theta};q)_\infty}{(\beta^{-1/2}te^{i\theta};q)_\infty(\beta^{-1/2}te^{-i\theta};q)_\infty},
\end{aligned}
$$

which, via analytic continuation, proves the conjecture when $x\in[-1,1]$ and $y=(\beta^{1/2}+\beta^{-1/2})/2$. Similarly, it follows that the conjecture is true whenever x or y is one of the points $\pm(\beta^{1/2}+\beta^{-1/2})/2$. Thus it remains to prove the conjecture when $x,y\in[-1,1]$. What seems to be needed is a basic analogue of (4.5) which can be applied to the ${}_8\phi_7$ series in (4.3).

R. Askey suggested that limit cases of the formulas for $K_t(x,y;\beta|q)$ and $L_t(x,y;\beta|q)$ in §2 could also be obtained as integral representations by using the q-gamma function

$$\Gamma_q(z)=\frac{(q;q)_\infty}{(q^z;q)_\infty}(1-q)^{1-z},$$

the q-integral

$$\int_0^1 f(u)\,d_qu=(1-q)\sum_{n=0}^{\infty}f(q^n)q^n$$

and the limits

$$\lim_{q\to1^-}\Gamma_q(z)=\Gamma(z),\qquad \lim_{q\to1^-}\int_0^1 f(u;q)\,d_qu=\int_0^1 f(u;1)\,du,$$

for suitable functions $f(u;q)$ where $f(u;1)=\lim_{q\to1^-}f(u;q)$. In particular, from (2.11)

(4.7)

$$
\begin{aligned}
K_t\big(x,y;q^\lambda|q\big)&=\frac{(1-t^2)\Gamma_q(2\lambda)}{\Gamma_q(\lambda)\Gamma_q(\lambda)}(1-2t\cos(\theta+\psi)+t^2)^{-1}\\
&\quad\cdot\int_0^1\frac{u^{\lambda-1}(uq;q)_\infty(t^2uq;q)_\infty(1-t^2u^2q^\lambda)}{(uq^\lambda;q)_\infty(t^2uq^\lambda;q)_\infty}\\
&\quad\cdot\left|\frac{(tuq^\lambda e^{i(\theta+\psi)};q)_\infty(tuq^{\lambda+1}e^{i(\theta-\psi)};q)_\infty}{(tuqe^{i(\theta+\psi)};q)_\infty(tue^{i(\theta-\psi)};q)_\infty}\right|^2 d_qu
\end{aligned}
$$

and on letting $q \to 1^-$ we obtain

$$\sum_{n=0}^{\infty} \frac{(n+\lambda)n!}{\lambda(2\lambda)_n} C_n^{\lambda}(x)C_n^{\lambda}(y)t^n \tag{4.8}$$

$$= \frac{(1-t^2)\Gamma(2\lambda)}{\Gamma^2(\lambda)(1-2t\cos(\theta+\psi)+t^2)}$$

$$\cdot \int_0^1 \frac{u^{\lambda-1}(1-u)^{\lambda-1}(1-t^2u)^{\lambda-1}(1-t^2u)\,du}{[1-2tu\cos(\theta+\psi)+t^2]^{\lambda-1}[1-2tu\cos(\theta-\psi)+t^2]^{\lambda+1}}, \qquad \lambda>0,$$

which is quite different from Watson's integral representation in [18, p. 292].

Similarly, it follows from (2.9) and (2.14), respectively, that

$$\sum_{n=0}^{\infty} \frac{(n+\lambda)n!}{\lambda(2\lambda)_n} C_n^{\lambda}(x)C_n^{\lambda}(y)t^n \tag{4.9}$$

$$= \frac{(1-t^2)\Gamma(2\lambda)}{\Gamma(\lambda+1)\Gamma(\lambda-1)} \int_0^1 \frac{u^{\lambda}(1-u)^{\lambda-2}(1-t^2u)^{\lambda-2}(1-t^2u^2)\,du}{[(1-2tu\cos(\theta+\psi)+t^2)(1-2tu\cos(\theta-\psi)+t^2)]^{\lambda}}$$

for $\lambda>1$, and

$$\sum_{n=0}^{\infty} \frac{n!}{(2\lambda)_n} C_n^{\lambda}(x)C_n^{\lambda}(y)t^n \tag{4.10}$$

$$= \frac{\Gamma(2\lambda)}{\Gamma^2(\lambda)} \int_0^1 \frac{u^{\lambda-1}(1-u)^{\lambda-1}(1-t^2u)^{\lambda-1}(1-t^2u^2)\,du}{[(1-2tu\cos(\theta+\psi)+t^2)(1-2tu\cos(\theta-\psi)+t^2)]^{\lambda}}$$

for $\lambda>0$.

Formulas (4.9) and (4.10) can also be proved directly by using the following rather strange looking limit cases of (2.8) and (2.13):

(4.11)

$${}_3F_2\left[\begin{matrix} a, 1+a/2, -b \\ a/2, a+b+1 \end{matrix}; t\right] = \frac{(1-t)\Gamma(a+b)}{\Gamma(a+1)\Gamma(b-1)} \int_0^1 u^a(1-u)^{b-2}(1-tu)^{b-2}(1-tu^2)\,du$$

when $|t|<1$, $a>-1$, $b>1$; and

$${}_2F_1\left[\begin{matrix} a, -b \\ a+b+1 \end{matrix}; t\right] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 u^{a-1}(1-u)^{b-1}(1-tu)^{b-1}(1-tu^2)\,du \tag{4.12}$$

when $|t|<1$ and $a, b>0$, which are easy to prove directly and have the important property that a and b are separated inside the integrals.

REFERENCES

[1] G. E. Andrews, *q-identities of Auluck, Carlitz and Rogers*, Duke Math. J., 33 (1966), pp. 575–581.

[2] R. Askey and M. E. H. Ismail, *A generalization of ultraspherical polynomials*, Studies in Pure Mathematics, P. Erdös, ed., Birkhauser, 1982.

[3] W. N. Bailey, *The generating function of Jacobi polynomials*, J. London Math. Soc., 13 (1938), pp. 8–12.

[4] D. M. Bressoud, *A simple proof of Mehler's formula for q-Hermite polynomials*, Indiana Univ. Math. J., 29 (1980), pp. 577–580.

[5] _____, *Linearization and related formulas for q-ultraspherical polynomials*, this Journal, 12 (1981), pp. 161–168.
[6] L. CARLITZ, *Some polynomials related to theta functions*, Ann. Math. Pura Appl. Ser. 4, 41 (1955), pp. 359–373.
[7] _____, *Some polynomials related to theta functions*, Duke Math. J., 24 (1957), pp. 521–527.
[8] G. GASPER, *Nonnegativity of a discrete Poisson kernel for the Hahn polynomials*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.
[9] MIZAN RAHMAN, *A product formula and a nonnegative Poisson kernel for Racah–Wilson polynomials*, Canad. J. Math., 22 (1980), pp. 1501–1517.
[10] _____, *Reproducing kernels and bilinear sums for q-Racah and q-Wilson polynomials*, Trans. Amer. Math. Soc., 273 (1982), pp. 483–508.
[11] L. J. ROGERS, *On a three-fold symmetry in the elements of Heine's series*, Proc. London Math. Soc., 24 (1893), pp. 171–179.
[12] _____, *On the expansion of some infinite products*, Proc. London Math. Soc., 24 (1893), pp. 337–352.
[13] _____, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–343.
[14] _____, *Third memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 26 (1895), pp. 15–32.
[15] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.
[16] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.
[17] G. N. WATSON, *Notes on generating functions of polynomials*: (2) *Hermite polynomials*, J. London Math. Soc., 8 (1933), pp. 194–199.
[18] _____, *Notes on generating functions of polynomials*: (3) *Polynomials and Legendre and Gegenbauer*, J. London Math. Soc., 81 (1933), pp. 289–292.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0001 $01.25/0

SOME ASYMPTOTIC ESTIMATES FOR HIGHER ORDER AVERAGING AND A COMPARISON WITH ITERATED AVERAGING*

JAMES A. MURDOCK†

Abstract. Asymptotic estimates for classical higher order averaging are obtained on intervals of length greater than $O(1/\varepsilon)$ when some of the averages vanish. These results are compared with results of Persek using iterated averaging, and the classical methods are found to be more powerful.

1. Introduction. We shall be concerned with the n-dimensional system of differential equations

$$\dot{x}=\varepsilon f(x,t,\varepsilon)=\varepsilon f_1(x,t)+\varepsilon^2 f_2(x,t)+\cdots+\varepsilon^n f_n(x,t)+\varepsilon^{n+1} r_{n+1}(x,t,\varepsilon) \tag{1}$$

where f is 2π-periodic in t and ε is a small positive real number. For such systems the traditional nth order averaging method, as described for instance in Perko [2], yields approximate solutions which retain accuracy $O(\varepsilon^n)$ on time intervals of length $O(1/\varepsilon)$. It is of interest whether one can "trade off" some of this accuracy for validity on a longer time interval; that is to say, we may ask whether the same approximate solution retains the decreased accuracy $O(\varepsilon^{n-j})$ on the expanded time interval of length $O(1/\varepsilon^{1+j})$ for certain integers j. Our first theorem (§2) asserts that this is true for $j=0,1,\cdots,l-1$ if the nth order averaged system corresponding to (1) begins with the term of order ε^l, in other words takes the form

$$\dot{z}=\varepsilon^l g_l(z)+\cdots+\varepsilon^n g_n(z). \tag{2}$$

The proof involves no new methods.

Persek [3] has defined a method which he calls "iterated averaging" which under certain conditions approximates system (1) by a system of the form

$$\dot{z}=\varepsilon^l h_l(z). \tag{3}$$

He then proves that solutions of (3) approximate those of (1) to order $O(\varepsilon)$ on an interval of length $O(1/\varepsilon^l)$. It is evident that (3) has the same form as (2) in the case $n=l$, although it is not clear a priori whether the function h_l constructed by Persek coincides with the classical g_l in this case. Nevertheless Persek's estimate of the accuracy of (3) coincides with our estimate for (2) if $n=l$ and if j is taken to be $l-1$. This prompts a comparison between g_l and h_l, which we carry out in §3 for the case $l=2$. Briefly the result is that g_2 and h_2 coincide when both are defined, but that the defineability of h_2 depends upon a hypothesis which is unnecessary under the traditional method. Thus Persek's result is (at least for the case $l=2$) a special case of ours (in §2). For $l>2$ it would be tedious to compare h_l and g_l but it is again apparent that g_l is always defined (and thus equation (2) exists provided that g_1 through g_{l-1} vanish), whereas h_l is defined only if h_1 through h_{l-1} vanish and in addition a further condition is satisfied, which does not correspond to any condition in the traditional theory.

*Received by the editors November 13, 1981.
†Department of Mathematics, Iowa State University, Ames, Iowa 50011.

2. Asymptotic estimates. It is shown in the classical theory of averaging that there exists a coordinate transformation

$$x=u(y,t,\varepsilon)=y+\varepsilon u_1(y,t)+\cdots+\varepsilon^n u_n(y,t) \tag{4}$$

2π-periodic in t and carrying (1) to

$$\dot{y}=\varepsilon g_1(y)+\cdots+\varepsilon^n g_n(y)+\varepsilon^{n+1}R_{n+1}(y,t,\varepsilon). \tag{5}$$

The transformation (4) is not unique and is usually normalized either by requiring that the average of each u_k vanish, or by requiring that each u_k vanish for $t=0$. The latter is called the stroboscopic method and has the advantage that (4) reduces to $x=y$ at $t=0$ and at all stroboscopic times $t=2\pi, 4\pi, \cdots$. Associated with (5) is the truncated system

$$\dot{z}=\varepsilon g_1(z)+\cdots+\varepsilon^n g_n(z) \tag{6}$$

which is wholly autonomous. Let $x(t,\varepsilon)$, $y(t,\varepsilon)$, $z(t,\varepsilon)$ denote solutions of (1), (5), and (6) defined in an interval $0\le\varepsilon\le\varepsilon_0$ whose initial conditions are related by $x(0,\varepsilon)=u(y(0,\varepsilon),0,\varepsilon)$ and $y(0,\varepsilon)=z(0,\varepsilon)$; note that in the stroboscopic case this reduces to $x(0,\varepsilon)=y(0,\varepsilon)=z(0,\varepsilon)$. The classical method of averaging proposes to construct from $z(t)$ an approximation to $x(t)$. Since $z(t)$ ought to be close to $y(t)$, and since $y(t)$ is related to $x(t)$ by (4), in view of the fact that (4) is assumed to hold at $t=0$, $x(t)$ should be well approximated by $u(z(t,\varepsilon),t,\varepsilon)$, an expression which is called the *improved nth approximation* to $x(t)$. It turns out, however, that there is no loss in asymptotic accuracy if the term of order n in (4) is omitted in forming the approximation (it must not be omitted, of course, in transforming (1) into (5)). Therefore the *nth approximation* to $x(t)$ is defined by

$$X(t,\varepsilon)=\hat{u}(z(t,\varepsilon),t,\varepsilon), \tag{7}$$

where

$$\hat{u}(z,t,\varepsilon)=z+\varepsilon u_1(z,t)+\cdots+\varepsilon^{n-1}u_{n-1}(z,t).$$

THEOREM. *Under the above hypotheses there exist positive constants c and c_0 such that $|x(t,\varepsilon)-X(t,\varepsilon)|\le c_0\varepsilon^n$ for $0\le t\le c/\varepsilon$ and $0\le\varepsilon\le\varepsilon_0$. Under the additional hypothesis that g_1 through g_{l-1} vanish, so that* (6) *takes the form* (2), *there exist for each $j=0,\cdots,l-1$ positive constants c_j such that $|x(t,\varepsilon)-X(t,\varepsilon)|\le c_j\varepsilon^{n-j}$ for $0\le t\le c/\varepsilon^{1+j}$.*

Proof. We prove the second assertion. The first, which is classical, is included by taking $l=1, j=0$.

Let K be a closed ball centered at $x(0,0)$ of radius R sufficiently large that $x(0,\varepsilon)$ and $y(0,\varepsilon)=z(0,\varepsilon)$ are contained in the concentric ball of radius $R/2$ for $0\le\varepsilon\le\varepsilon_0$. Since K is compact there exists in view of (1), (2) and (5) a constant M such that $d|y|/dt$ and $d|z|/dt$ are less than $M\varepsilon^l$ for $0\le\varepsilon\le\varepsilon_0$ as long as y and z remain in K. Since $|y|$ and $|z|$ must drift by at least $R/2$ from their initial positions in order to leave K, and since in time t they can drift at most $M\varepsilon^l t$ while in K, it follows that there exists a constant $c>0$ such that y and z remain in K for $0\le t\le c/\varepsilon^l$. Hence on this interval (the longest considered in the theorem) one may use, for all functions of y and z, their upper bounds and Lipschitz constants on K.

Letting $\rho=|y(t,\varepsilon)-z(t,\varepsilon)|$ one immediately finds from (2) and (5) Lipschitz constants $L_l,\cdots,L_n$ and a bound B such that $d\rho/dt\le(\varepsilon^l L_l+\cdots+\varepsilon^n L_n)\rho+\varepsilon^{n+1}B$. Solving this linear differential inequality with initial condition $\rho=0$ yields $\rho\le\varepsilon^{n+1}B\delta^{-1}(e^{\delta t}-1)$ where $\delta=\varepsilon^l L_l+\cdots+\varepsilon^n L_n$. Estimating the factor $e^{\delta t}-1$ requires some care. The tempting answer $e^{\delta t}-1=O(\delta t)=O(\varepsilon^l t)$ is correct for the time intervals with which we

are concerned, but this requires proof. Namely $e^x-1=x(1+x/2!+\cdots)=x\phi(x)$ and hence for x in any bounded interval there is a constant k such that $e^x-1\le kx$; if $x=\delta t$ and $0\le t\le c/\varepsilon^l$ and $0<\varepsilon\le\varepsilon_0$ then x is bounded and $e^{\delta t}-1\le k\delta t$. (On longer intervals $e^{\delta t}-1$ can approach infinity faster than δt.) Since $\delta^{-1}=O(1/\varepsilon^l)$ we have $\rho=O(\varepsilon^{n+1}t)$ as long as y and z remain in K, that is, at least on the interval $0\le t\le c/\varepsilon^l$. It follows that on any interval $0\le t\le c/\varepsilon^{1+j}$, $j=0,\cdots,l-1$, one has $|y(t,\varepsilon)-z(t,\varepsilon)|=\rho=O(\varepsilon^{n-j})$; that is, this quantity is bounded by a constant times ε^{n-j} for $0\le t\le c/\varepsilon^{1+j}$ and $0\le\varepsilon\le\varepsilon_0$. Now using the Lipschitz constant for u on K one finds $|x(t,\varepsilon)-u(z(t,\varepsilon),t,\varepsilon)|=|u(y(t,\varepsilon),t,\varepsilon)-u(z(t,\varepsilon),t,\varepsilon)|=O(\varepsilon^{n-j})$ on the same interval. But

$$|u(z(t,\varepsilon),t,\varepsilon)-X(t,\varepsilon)|=|u(z(t,\varepsilon),t,\varepsilon)-\hat{u}(z(t,\varepsilon),t,\varepsilon)=O(\varepsilon^n)$$

for all time; adding the last two estimates proves the theorem. Q.E.D.

In the proof it is seen that the final term in the transformation u is unnecessary in constructing $X(t)$ because the error committed by leaving it out is of the same order as the error already present. By the same reasoning we see that for $j>0$, where the possible accuracy is at most $O(\varepsilon^{n-j})$, we may omit j additional terms from u in forming X. In particular, in the case $n=l$, $j=l-1$, it is not necessary to use u at all and we obtain

COROLLARY. *When* (2) *reduces to* $\dot{z}=\varepsilon^l g_l(z)$, *there exist positive constants* c *and* c_0 *such that* $|x(t,\varepsilon)-z(t,\varepsilon)|<c_0\varepsilon$ *for* $0\le t\le c/\varepsilon^l$.

This form of the theorem is the one most directly comparable to the work of Persek. The comparison is carried out in the next section.

3. Comparison of two averaging methods. In order to calculate $g_2(z)$ it is necessary to recall how (4) and (5) are constructed. It is clear a priori that any transformation of the form (4) carries (1) into a system of the form

$$\dot{y}=\varepsilon g_1(y,t)+\cdots+\varepsilon^n g_n(y,t)+\varepsilon^{n+1}R_{n+1}(y,t,\varepsilon). \tag{8}$$

From (1), (4), and (8) one calculates that the f's, u's, and g's are related by

$$\begin{aligned}\frac{\partial u_1}{\partial t}(y,t)&=f_1(y,t)-g_1(y,t),\\ \frac{\partial u_2}{\partial t}(y,t)&=\left\{f_2+\frac{\partial f_1}{\partial y}u_1-\frac{\partial u_1}{\partial y}g_1\right\}_{(y,t)}+g_2(y,t)\end{aligned} \tag{9}$$

with similar equations for the higher u_n's. (Briefly, to obtain (9) differentiate (4), insert (5) and compare this with the result of inserting (4) into (1).) It is clear that (9) admits solutions for u_1 and u_2 which are periodic in t if and only if the right-hand sides have zero mean value. Now to achieve (5) the g_i must be independent of t; thus (9) dictates that $g_1(y)$ must be the average of $f_1(y,t)$ and $g_2(y)$ must be the average of the expression in braces.

We wish to consider the case in which the averaged system takes the form (2) with $l=2$. Thus we now assume $g_1(y)=0$, which is to say that the average of $f_1(y,t)$ vanishes. In this case $u_1(y,t)=\int_a^t f_1(y,s)\,ds$ with a arbitrary ($a=0$ gives the stroboscopic method). Inserting this into the expression in braces and averaging gives the following formula for g_2, in which the dependence upon the choice of a is made explicit:

$$g_2(y,a)=\frac{1}{2\pi}\int_0^{2\pi}\left\{f_2(y,t)+\frac{\partial f_1}{\partial y}(y,t)\int_a^t f_1(y,s)\,ds\right\}dt. \tag{10}$$

Persek's function $h_2(z)$ is defined as follows in our notation (compare his $\bar{E}^{(2)}$, [3, p. 416]). First assume the average of $f_1(y,t)$ vanishes, as we have done above. Next define

$$H_2(z,\tau)=\frac{1}{2\pi}\int_{\tau}^{\tau+2\pi}\left\{f_2(z,t)+\frac{\partial f_1}{\partial y}(y,t)\int_{\tau}^{t}f_1(y,s)\,ds\right\}dt. \tag{11}$$

If the latter expression is independent of τ, it is defined to be $h_2(z)$; if it is not independent of τ, $h_2(z)$ is not defined.

Now it is clear that the bracketed expression in (10) is periodic in t, since we have assumed that f_1 has zero mean value. Therefore $\int_0^{2\pi}$ in (10) may be replaced by $\int_{\tau}^{\tau+2\pi}$ for any τ. The only remaining difference between (10) and (11) is that a in (10) is replaced by τ in (11). Thus we see that

$$H_2(z,\tau)=g_2(z,\tau). \tag{12}$$

Thus the sole difference between Persek's average (for $n=l=2$) and ours is that Persek must assume (11) independent of τ, whereas the corresponding quantity a in (10) enters as an arbitrary constant and requires no additional assumptions. It is clear that the assumption that (11) is independent of τ is a very strong assumption and one which gains no advantage.

With regard to higher order terms the following situation obtains. Our g_k is always defineable, and is not unique but rather depends upon the choices of integration constants in solving for u. On the other hand h_k is only defined if two conditions are met: $h_1,\cdots,h_{k-1}$ must be defined and vanish, and a certain function $H_k(z,\tau)$ (which Persek calls $\bar{E}^{(k)}$) must be independent of τ. The presence of the latter restriction indicates that there are likely to be many cases in which (2) takes the form $\dot{z}=\epsilon^l g_l(z)$ and yet (3) cannot be formulated. It seems reasonable to conjecture (based on the case $l=2$) that h_l exists precisely when g_l is unique (i.e., independent of the choices made in u) and that in this case $h_l=g_l$. Proof of such a theorem, if true, would involve notational difficulties but might be attempted (if it were considered important) by the use of the operators constructed by Musen [1] for use in averaging methods, based upon a formula of St. Faa de Bruno.

REFERENCES

[1] P. Musen, *On the high order effects in the methods of Krylov–Bogoliubov and Poincaré*, J. Astronaut. Sci. 12 (1965), pp. 129–134.

[2] L. M. Perko, *Higher order averaging and related methods for perturbed periodic and quasi-periodic systems*, SIAM J. Appl. Math., 17 (1968), pp. 698–723.

[3] S. C. Persek, *Hierarchies of iterated averages for systems of ordinary differential equations with a small parameter*, SIAM J. Math. Anal., 12 (1981), pp. 413–420.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0002 $01.25/0

A NONLINEAR PROBLEM ARISING FROM COMBUSTION THEORY: LIÑÁN'S PROBLEM*

S. P. HASTINGS† AND A. B. POORE‡

Abstract. The differential equation $y''=y\exp(\alpha x-y)$, $\lim_{x\to-\infty}\frac{dy}{dx}=-\theta$ and $\lim_{x\to\infty}\frac{dy}{dx}=0$ for $\theta>0$ governs the thin reaction diffusion zone in many diverse problems in combustion theory. This problem with $\theta=1$ is known as Liñán's problem and continues to arise through the use of large activation energy asymptotics in the study of various combustion phenomena. The main issues for this problem are those of existence and uniqueness which we establish for each positive α and θ.

1. Introduction. Activation energy asymptotics has been firmly established as an effective analytical technique for dealing with the Arrhenius rate function $\exp(-\gamma/T)$ (γ is the activation energy and T, the temperature) which is present in many partial differential equations which govern chemical processes in combustion and chemical reactor theory. The idea of using large activation energies and matched asymptotics on combustion problems was popularized by F. A. Williams [11], although the paper of W. B. Bush and F. E. Fendell [3] appeared earlier. An explanation of the method and applications to many diverse problems can be found in the book by J. Buckmaster and G. S. S. Ludford [2] or in the extensive literature of which a small representation is given in the references [1]–[5], [7]–[11].

In a typical application the use of large activation energy asymptotics renders an intractable nonlinear problem tractable in certain regions of the problem, whereas the solution over the entire region of consideration is obtained by connecting the "outer" solutions by using matched asymptotics. Fundamental to this matching is, in many cases, a nonlinear differential equation which governs the (thin) reaction-diffusion zone (internal or boundary layer) and whose solution is critical to the overall analysis of the combustion phenomenon under investigation. Although many such problems arise, a problem which has a demonstrated permanence and continues to arise is Liñán's problem. It first appeared in 1974 in Liñán's paper on counterflow diffusion flames [8]. Since that time, this problem and minor variations have been found to govern the (thin) reaction-diffusion zone in such problems as the burning of monopropellant drops, detonations and fast deflagration waves [9], the flame-front region problem studied by W. B. Bush and S. F. Fink [4] and the nonadiabatic tubular reactor [7].

A slight generalization of Liñán's problem as it arises from matching is to establish the existence and uniqueness of a solution of

$$\frac{d^2y}{dx^2}=\frac{1}{2}ye^{\alpha x-y},\quad \lim_{x\to+\infty}\frac{dy}{dx}=0,\quad \lim_{x\to-\infty}\frac{dy}{dx}=-\theta, \tag{1}$$

where $\theta>0$. Liñán's problem corresponds to $\theta=1$. A second formulation of this problem which is more appropriate for the flame-front region flow problem studied by W. B. Bush and S. F. Fink [4] is

$$\frac{d^2y}{dx^2}=\frac{1}{2}ye^{\alpha x-y},\quad \lim_{x\to+\infty}y=0,\quad \lim_{x\to-\infty}\frac{dy}{dx}=-\theta \tag{2}$$

for positive α and θ.

* Received by the editors December 2, 1981 and in revised form March 15, 1982.
† Department of Mathematics, State University of New York at Buffalo, Buffalo, New York 14222.
‡ Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.

The goal then of this work is to establish the existence and uniqueness for both problems (1) and (2) for all positive values of α and θ. In fact, the two solutions are the same. The outline of the argument is given in the next section followed by the proofs in §3.

2. Outline of the argument. Let $y(x; x_0, \beta)$ denote the solution of the initial value problem

$$y'' = \frac{1}{2} y e^{\alpha x - y}, \quad y(x_0) = 1, \quad y'(x_0) = \beta, \tag{3}$$

where $\alpha > 0$. (The initial condition $y(x_0) = 1$ could be replaced by $y(x_0) = y_0$ for any $y_0 \in (0, 1]$.) For each fixed x_0 we first prove that there exists a β such that $y'(x; x_0, \beta) \to 0$ as $x \to +\infty$. This is accomplished by the first showing that the two sets

$$R_1 = \{\beta : y(x; x_0, \beta) = 0 \text{ for some } x > x_0\}$$

and

$$R_2 = \{\beta : y'(x; x_0, \beta) > 0 \text{ for some } x > x_0\}$$

are nonempty, open and disjoint subsets of $\mathbb{R}$, the reals. By connectedness of the reals, there is a β in the complement of $R_1 \cup R_2$. The corresponding solution $y(x; x_0, \beta)$ is then shown to have the desired decay properties as $x \to +\infty$. The existence and uniqueness of this solution is the content of

THEOREM 1. *For each $\alpha > 0$ and arbitrary x_0, there exists a unique solution $y(x; x_0)$ of*

$$y'' = \frac{1}{2} y e^{\alpha x - y}, \quad y(x_0) = 1, \quad y' \to 0 \quad \text{as } x \to +\infty. \tag{4}$$

Furthermore, $y \to 0$ as $x \to +\infty$.

To obtain the existence of a solution of problem (1), we let $y(x, x_0)$ denote the unique solution in Theorem 1 and define

$$L_1 = \{x_0 : y'(x; x_0) < -\theta \text{ for some } x < x_0\}$$

and

$$L_2 = \left\{x_0 : y'(x; x_0) > -\theta + \frac{1}{2\alpha} e^{\alpha x} \text{ for some } x < x_0\right\}.$$

Again these sets are shown to be nonempty, open and disjoint. The x_0 in the complement of $L_1 \cup L_2$ yields a solution of (1). Thus, we have

THEOREM 2. *Let $y(x; x_0)$ denote the unique solution of problem* (4). *Then for each $\alpha > 0$ there is an x_0 such that $\lim_{x \to +\infty} y'(x; x_0) = -\theta$ so that this $y(x; x_0)$ is a solution of problem* (1).

Given existence we next establish

THEOREM 3. *For each $\alpha > 0$ and $\theta > 0$ the solution of problem* (1) *is unique and $\lim_{x \to +\infty} y = 0$.*

We observe that Theorem 3 implies that the unique solution of problem (1) is in fact a solution of problem (2). What is more, if y is a solution of problem (2) for positive α and θ, then $0 = \lim_{x \to +\infty} y(x) = y(x_0) + \int_{x_0}^{\infty} dy/d\xi \, d\xi$ implies $\lim_{x \to +\infty} dy/dx = 0$. Thus we have the equivalence of the two solutions. This observation is stated in

THEOREM 4. *For positive α and θ the unique solution of* (1) *is the unique solution of* (2).

The proofs of Theorems 1, 2 and 3 are given in the next section.

3. Proofs. The first goal is to show that the sets R_1 and R_2 are nonempty, open and disjoint subsets of $\mathbb{R}$. That R_1 is nonempty is contained in

LEMMA 1. *For each $\alpha>0$ there exists a B such that for every $\beta\leq B$, the solution of* (3) *has a zero which is greater than x_0. Thus the set R_1 is nonempty.*

Proof. First observe that $ye^{-y}<1$ for all real y and consider the problem $z''=\frac{1}{2}e^{\alpha x}$, $z(x_0)=1$ and $z'(x_0)=\gamma$ which is uniquely solvable by

$$z(x)=\frac{1}{2\alpha^2}\{e^{\alpha x}-e^{\alpha x_0}\}+1+\left\{\gamma-\frac{1}{2\alpha}e^{\alpha x_0}\right\}(x-x_0).$$

Let $\gamma^*(x_0)$ be the unique value of γ for which $z(x)$ has a single zero. Then $\gamma>\gamma^*(x_0)$ implies that z has no zeros and $\gamma<\gamma^*(x_0)$, that z has two zeros (note that $\gamma^*(x_0)$ is negative and is defined by $z(x_1)=z'(x_1)=0$, where x_1 is the single zero). For $\beta\leq\gamma^*(x_0)$ consider the two problems

$$\begin{aligned} z''&=\frac{1}{2}e^{\alpha x}, & y''&=\frac{1}{2}ye^{\alpha x-y},\\ z(x_0)&=1, & y(x_0)&=1,\\ z'(x_0)&=\beta, & y'(x_0)&=\beta. \end{aligned}$$

If $d(x)=z(x)-y(x)$, then $d''=\frac{1}{2}e^{\alpha x}-\frac{1}{2}ye^{\alpha x-y}\geq 0$ and $d(x_0)=d'(x_0)=0$. Thus $d(x)\geq 0$ or $z(x)\geq y(x)$ as long as $y(x)$ exists. Let x_1 be the first zero of $z(x)$, which exists since $\beta\leq\gamma^*(x_0)$. If the solution $y(x)$ exists on $[x_0,x_1]$, then y must have a zero at some point of $[x_0,x_1]$. If the maximal interval of existence of y for $x>x_0$ is $[x_0,b)$ where $b\leq x_0$, then $y\to+\infty$ or $y\to-\infty$ as $x\to b-0$ since $\frac{1}{2}ye^{\alpha x-y}$ is continuous everywhere [5, p. 17]. But y is bounded above on $[x_0,b)$ by $z(x)$ so that $y\to-\infty$ which implies the existence of the zero. Let $B=\gamma^*(x_0)$ in the statement of the lemma.Q.E.D.

We next show that any solution of (3) with $\beta\in R_1$ is strictly decreasing.

LEMMA 2. *Let $\beta\in R_1$. Then $dy/dx(x;x_0,\beta)<0$ for all $x\geq x_0$ for which the solution exists. Thus the zero of this solution is unique.*

Proof. Any solution of $y''=\frac{1}{2}ye^{\alpha x-y}$ satisfies $y''y>0$ except at $y=0$. Hence y cannot have a positive maximum, a negative minimum or an inflection point except at $y=0$. In particular, if y vanishes, say initially at $x_1>x_0$, then $y'<0$ on $[x_0,x_1)$. Also, $y'(x_1)<0$, for if $y=y'=0$ at x_1, then $y\equiv 0$ by uniqueness. Thus $y'<0$ for x just to the right of x_1 and must remain negative thereafter since y can have no negative minimum. Q.E.D.

With these two lemmas we now prove

LEMMA 3. *The sets R_1 and R_2 are nonempty, disjoint and open subsets of $\mathbb{R}$, the reals.*

Proof. Lemma 1 implies R_1 is nonempty and any $\beta\geq 0$ is in R_2. Lemma 2 implies that any solution $y(x;x_0,\beta)$ of (3) with $\beta\in R_1$ cannot have a positive slope which implies the disjointness of R_1 and R_2. That R_2 is open follows from the continuity of the solution with respect to the initial data and the strict inequalities. To show that R_1 is open, observe from the proof of Lemma 2 that $\beta\in R_1$ is and only if y becomes strictly negative. This implies the result. Q.E.D.

We now proceed with the proof of Theorem 1. Since R_1 and R_2 are nonempty, disjoint and open subsets of the reals, connectedness of the reals implies the existence of a β, say β^*, in the complement of $R_1\cup R_2$. Let y^* denote the corresponding solution. Since $\beta^*\notin R_1\cup R_2$, $0<y^*<1$ and $\frac{dy^*}{dx}<0$ for all $x\geq x_0$ for which y^* exists. But the boundedness of y^* implies [5, p. 17] that y^* exists on $[x_0,+\infty)$. Now $\lim_{x\to\infty}y^*$ exists and is nonnegative. If it is positive, $\frac{dy^*}{dx}(x)=\beta^*+\int_{x_0}^{x}\frac{1}{2}y^*e^{\alpha x-y^*}\,d\xi$ becomes positive as

$x \to +\infty$, a contradiction to $\beta^* \notin R_2$. Thus $\lim_{x\to\infty} y^* = 0$. Also, $\frac{d^2y^*}{dx^2} > 0$ and $\frac{dy^*}{dx} < 0$ implies $\lim_{x\to\infty} \frac{dy^*}{dx}$ exists. If it is negative, then

$$y^* - 1 = \int_{x_0}^{x} \frac{dy^*}{d\xi} d\xi \to -\infty \quad \text{as } x \to +\infty,$$

so that $\beta^* \in R_1$, a contradiction. Thus y^* and $\frac{dy^*}{dx}$ tend to zero as $x \to +\infty$.

To prove uniqueness and thus complete the proof of Theorem 1, suppose problem (4) has two solutions, y_1 and y_2 with $y_1'(x_0) > y_2'(x_0)$. Since $0 < y_i < 1$ on $[x_0, \infty)$ and $\frac{d}{dy}(ye^{-y}) > 0$ for $0 < y < 1$, it follows that $y_1'' > y_2''$ as long as $y_1 > y_2$. Since $y_1' > y_2'$ initially (at x_0), it is clear that $y_1 - y_2$ and $y_1' - y_2'$ are positive and increasing on (x_0, ∞), which contradicts the assumption that y_1' and y_2' tend to zero as $x \to +\infty$. This completes the proof of Theorem 1.

The next goal is to show that the sets L_1 and L_2 are nonempty, disjoint and open subsets of the reals. To establish openness we shall make use of:

LEMMA 4. *Let $y(x; x_0)$ denote the unique solution of problem* (4) *as given in Theorem* 1. *Then $y'(x_0; x_0) = \frac{\partial y}{\partial x}(x; x_0)|_{x=x_0}$ depends continuously on x_0.*

Proof. If $y'(x_0; x_0)$ does not depend continuously on x_0, then there is a sequence $\{n_j\}$ converging to some $\hat{x}_0$ such that $\{y'(n_j; n_j)\}$ does not converge or has a limit distinct from $y'(\hat{x}_0; \hat{x}_0)$. Let $\beta_j = y'(n_j; n_j)$. We next show that $\{\beta_j\}$ is a bounded sequence.

For each x_0 let $\gamma^*(x_0)$ denote the number defined in the proof of Lemma 1. Note that $\gamma^*(x_0) < 0$ and $\gamma^*(x_0)$ is continuous for all x_0. Also, from the definition of $\gamma^*(x_0)$ and the comparison argument of Lemma 1, $\gamma^*(n_j) \le \beta_j < 0$. Continuity of $\gamma^*(x_0)$ and $n_j \to \hat{x}_0$ implies $\{\beta_j\}$ is bounded. Hence some subsequence $\{\beta_{j_k}\}$ converges to $\hat{\beta} \neq y'(\hat{x}_0; \hat{x}_0)$ by our assumption that Lemma 4 is false; otherwise, $\limsup \beta_j = \liminf \beta_j = y'(\hat{x}_0, \hat{x}_0)$. Let $y(x; x_0, \hat{\beta})$ denote the solution of (3) with $\beta = \hat{\beta}$. We now show $\lim_{x\to+\infty} y'(x; x_0, \hat{\beta}) = 0$.

If not, then either $y'(x_1; x_0, \hat{\beta}) > 0$ for some $x_1 > x_0$ or $y(x_2; x_0, \hat{\beta}) < 0$ for some $x_2 > x_0$. Consider the first case. For sufficiently large k, $y'(x_1; n_{j_k}) > 0$ as well since $n_{j_k} \to \hat{x}_0$. But then $\lim_{x\to+\infty} y'(x_1; n_{j_k}) \neq 0$, a contradiction. A similar absurdity occurs if $y(x_2; x_0, \hat{\beta}) < 0$.

Thus we have two distinct solutions of problem (4) at $x_0 = \hat{x}_0$, namely $y(x; \hat{x}_0)$ and $y(x; x_0, \hat{\beta})$. But this contradicts Theorem 1. Hence $y'(x_0; x_0)$ must depend continuously on x_0. Q.E.D.

We now establish

LEMMA 5. *The sets L_1 and L_2 are nonempty, disjoint and open sets of the reals.*

Proof. To show that L_1 is not empty, we show $y'(x_0; x_0) \to -\infty$ as $x_0 \to +\infty$. If this is not the case, then there is an $M > 0$ and an arbitrarily large x_0 with $y'(x_0; x_0) \ge -M$. Since $y'' > 0$, $y \ge 1 - M(x - x_0)$ on, say, $x_0 \le x \le x_0 + \frac{1}{2M}$. Hence $y \ge \frac{1}{2}$ and $y'' > \frac{1}{4}\exp(-\frac{1}{2} + \alpha x)$ on $[x_0, x_0 + \frac{1}{2M}]$. Integrating this shows that for large x_0, $y'(x_0 + \frac{1}{2M}; x_0) \ge 0$, which is impossible. Thus L_1 is not empty. A similar argument shows that $y'(x_0; x_0) \to 0$ as $x_0 \to -\infty$, so that $y'(x_0; x_0) - \frac{1}{2\alpha} e^{\alpha x_0} > -\theta$ for large negative x_0. For such an x_0 continuity implies $y'(x; x_0) > -\theta + \frac{1}{2\alpha} e^{\alpha x}$ for some $x < x_0$. Hence L_2 is not empty.

To establish the disjointness of L_1 and L_2 we first observe that if $x_0 \in L_1$, then $y'(x; x_0) < -\theta$ for all $x \le x_1 < x_0$ for some x_1. If $x_0 \in L_2$, $y'(x; x_0) - \frac{1}{2\alpha} e^{\alpha x} > -\theta$ for some $x_2 < x_0$. But $\frac{d}{dx}(y'(x; x_0) - \frac{1}{2\alpha} e^{\alpha x}) \le 0$ for all x implies $y'(x; x_0) - \frac{1}{2\alpha} e^{\alpha x}$ decreases as x increases or increases as x decreases. Thus, $y'(x; x_0) - \frac{1}{2\alpha} e^{\alpha x} > -\theta$ for all $x \le x_2$, and x_0 cannot be in both L_1 and L_2.

That L_1 and L_2 are open follows from the continuity of $y'(x_0; x_0)$ with respect to x_0 (Lemma 4) and the strict inequalities used in the definition of L_1 and L_2. Q.E.D.

Given Lemma 5, connectedness of the reals implies the existence of $x_0 \in \mathbb{R} - (R_1 \cup R_2)$ and a corresponding solution $y(x; x_0)$ as described in Theorem 1. Now $y'(x; x_0) \geq -\theta$ since $x_0 \notin L_1$ and $y'(x; x_0) \leq -\theta + \frac{1}{2\alpha} e^{\alpha x}$ since $x_0 \notin L_2$ for all $x \leq x_0$ for which the solution exists. Thus $1 - \theta(x - x_0) \leq y(x; x_0) \leq 1 - \theta(x - x_0) + \frac{1}{2}(e^{\alpha x} - e^{\alpha x_0})$ so that y is bounded on every finite interval. Thus, since the differential equation is continuous for all x and y, this solution may be continued [5, p. 17] to $(-\infty, \infty)$. Furthermore, $-\theta \leq y'(x; x_0) \leq -\theta + \frac{1}{2\alpha} e^{\alpha x}$ implies $\lim_{y \to -\infty} y'(x; x_0) = -\theta$ for $\alpha > 0$. Therefore, this $y(x; x_0)$ is a solution to problem (1). Having thus completed the proof of Theorem 2, we turn to uniqueness.

If y is a solution of problem (1), then $y > 0$ and $y' < 0$ on $(-\infty, \infty)$ and $y \to 0$ as $x \to +\infty$. Suppose there are two solutions y_1 and y_2. We can assume that $y_1(0) < y_2(0)$ and if $y_1(0) = y_2(0)$ then $y_1'(0) < y_2'(0)$. Hence we can assume $y_1 < y_2$ on some interval $(0, \varepsilon)$.

Let

$$Q_i(x) = \frac{(y_i'(x))^2}{2} - K(y_i(x))e^{\alpha x} \quad \text{for } i = 1 \text{ and } 2,$$

where

$$K(y) = \int_0^y k(u)\,du \quad \text{and} \quad k(u) = \frac{1}{2} u e^{-u}.$$

Then $Q_i'(x) = y_i'[y_i'' - k(y_i)e^{\alpha x}] - \alpha K(y_i)e^{\alpha x}$ so that

$$Q_i'(x) = -\alpha K(y_i)e^{\alpha x} < 0. \tag{5}$$

Next we establish

LEMMA 6. $Q_1(0) \leq Q_2(0)$.

Proof. Suppose $Q_1(0) > Q_2(0)$. Since $k(u) > 0$ for $u > 0$, $\frac{d}{dy} K(y) > 0$ for $y > 0$ and from (5), $Q_1' > Q_2'$ as long as $y_1 < y_2$ and, in particular, on some interval $(0, \varepsilon)$. Suppose there is some smallest $x_0 > 0$, where $y_1(x_0) = y_2(x_0)$. Then $Q_1 > Q_2$ and $Q_1' > Q_2'$ on $(0, x_0)$. Hence $Q_1(x_0) > Q_2(x_0)$. Since $y_1'(x_0)$ and $y_2'(x_0)$ are negative, $y_1'(x_0) < y_2'(x_0)$, so $y_1 > y_2$ on some interval $(x_0 - \delta, x_0)$, which contradicts the definition of x_0. Hence $y_1 < y_2$ on $(0, \infty)$. But this implies $Q_1 > Q_2$ and $Q_1' > Q_2'$ on $(0, \infty)$. To obtain a contradiction we show that Q_1 and Q_2 must both tend to 0 as $x \to \infty$. Let y denote either y_1 or y_2 and Q, the corresponding Q_1 or Q_2. Now y and y' approach zero and we must show that $K(y)e^{\alpha x} \to 0$. Using Hospital's rule we find that it is sufficient to prove that $ye^{-y}e^{\alpha x} \to 0$; i.e., that $y'' \to 0$. Clearly $\lim_{x \to \infty} \inf y'' = 0$ for otherwise $y' \to +\infty$. Thus, if the result is false, there is an $\varepsilon > 0$ and a sequence $\{x_n\}$ of local maxima of y'' such that $x_n \to \infty$ and $y''(x_n) \geq \varepsilon$. By differentiating $y'' = \frac{1}{2} y e^{-y + \alpha x}$ twice and using $y \to 0$, $y' \to 0$, we see that $y^{(iv)}(x_n) > 0$ for large n, so that x_n could not be local maxima of y''. This proves the lemma. Q.E.D.

Now let $v_i(x) = y_i(-x)$ and $P_i(x) = Q_i(-x)$ and consider $x \in [0, \infty)$. Then $P_i'(x) = -Q_i'(-x) = \alpha K(v_i(x)) \exp(-\alpha x)$. Also, by Lemma 6, $P_1(0) \leq P_2(0)$. Furthermore, $v_1 > v_2$ if and only if $P_1' > P_2'$. Finally, $v_1(0) \leq v_2(0)$.

If $v_1(0) = v_2(0)$ ($y_1(0) = y_2(0)$), then $v_1'(0) > v_2'(0) > 0$ since we assumed $y_1'(0) < y_2'(0)$. However, this contradicts $P_1(0) \leq P_2(0)$, so $P_1 < P_2$ on some interval $(0, \varepsilon)$. Furthermore, $P_1' < P_2'$ as long as $v_1 < v_2$.

Suppose there is a first $x_0>0$ with $v_1(x_0)=v_2(x_0)$. Then $P_1(x_0)<P_2(x_0)$ implies $v_1'(x_0)<v_2'(x_0)$, contradicting $v_1<v_2$ on $(0,x_0)$. Hence $v_1<v_2$, $P_1<P_2$ and $P_1'<P_2'$ on $(0,\infty)$. Since $-y_i'(-x)=v_i'(x)\to-\theta$ as $x\to+\infty$, it must eventually be the case that $-K(v_1)<-K(v_2)$ or $v_1>v_2$, again a contradiction. The assumption that there are two solutions is untenable, and the uniqueness is established.

Acknowledgment. The second author wishes to express his appreciation to K. Kirchgassner for his stimulating and insightful discussions of this problem.

REFERENCES

[1] J. W. Bebernes and D. R. Kassoy, *A mathematical analysis of blowup for thermal reactions—the spatially nonhomogeneous case*, SIAM J. Appl. Math., 40 (1981), pp. 476–484.

[2] J. Buckmaster and G. S. S. Ludford, *Theory of Laminar Flames*, Cambridge Univ. Press, Cambridge, 1982.

[3] W. B. Bush and F. E. Fendell, *Asymptotic analysis of laminar flame propagation for general Lewis number*, Comb. Sci. Tech., 1 (1970), pp. 421–428.

[4] W. B. Bush and S. F. Fink, *Planar premixed-flame/end-wall interaction: The jump conditions across the thin flame*, Quart Appl. Math., 38 (1981), pp. 427–438.

[5] J. K. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[6] A. K. Kapila and B. J. Matkowsky, *Reactive-diffusive systems with Arrhenius kinetics: multiple solutions, ignition and extinction*, SIAM J. Appl. Math., 36 (1979), pp. 373–389.

[7] A. K. Kapila and A. B. Poore, *The steady response of a nonadiabatic tubular reactor: new multiplicities*, Chem. Engrg. Sci., 37 (1982), pp. 57–68.

[8] A. Liñán, *The asymptotic structure of counterflow diffusion flames for large activation energy*, Acta Astronaut, 1 (1974), pp. 1007–1039.

[9] G. S. S. Ludford and D. S. Stewart, *Mathematical questions from combustion theory*, Trans. Twenty-Sixth Conference of Army Mathematicians, USARO Report 81-1, January, 1981, pp. 53–66.

[10] B. J. Matkowsky and G. I. Sivashinsky, *Acceleration effects on the stability of flame propagation*, SIAM J. Appl. Math., 37 (1979), pp. 669–685.

[11] F. A. Williams, *Theory of combustion in laminar flows*, Ann. Rev. Fluid Mech., 3 (1971), pp. 171–189.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0003 $01.25/0

NONLINEAR EIGENVALUE PROBLEMS ON INFINITE INTERVALS*

PETER A. MARKOWICH† AND RICHARD WEISS‡

Abstract. This paper is concerned with nonlinear eigenvalue problems of boundary value problems for ordinary differential equations posed on an infinite interval. It is shown that—under certain analyticity assumptions—a domain in the complex plain can be identified, in which all eigenvalues are isolated. A common way to solve such problems is to cut the infinite interval at a finite point and to impose additional, so-called asymptotic boundary conditions at this far end. The eigenvalue problem on the finite interval obtained this way can be solved by an appropriate code. In this paper suitable asymptotic boundary conditions are devised and the order of convergence, as the length of the finite interval converges to infinity, is investigated. Exponential convergence is shown for well posed approximating problems.

AMS-MOS subject classification (1980). Primary 34B25, 34D05, 34B05, 34P30

Key words. spectral theory of boundary value problems, asymptotic properties, asymptotic expansion, nonlinear eigenvalue problems

1. Introduction. This paper is concerned with nonlinear eigenvalue problems of the form

$$y' = t^{\alpha} A(t,\lambda) y, \qquad 1 \le t < \infty, \quad \alpha > -1, \tag{1.1}$$

$$B(\lambda) y(1) = 0, \tag{1.2}$$

$$y \in C([1,\infty]) : \Leftrightarrow y \in C([1,\infty)) \quad \text{and} \quad \lim_{t\to\infty} y(t) \text{ exists} \tag{1.3}$$

where y is an n-vector and $A(t,\lambda)$ is an $n \times n$ matrix. Equation (1.1) has a singularity of the second kind of rank $\alpha+1$ at $t=\infty$.

A solution of (1.1), (1.2), (1.3) is given by a pair (μ, y), $\mu \in \mathbb{C}$ such that $y \not\equiv 0$ satisfies (1.1), (1.2) with $\lambda = \mu$ and (1.3). Eigenvalue problems on infinite intervals occur frequently in quantum mechanics and in fluid mechanics, when the stability of laminar flows over infinite media is investigated (see Ng and Reid (1980)).

De Hoog and Weiss (1980a) and Markowich (1982a) treated linear eigenvalue problems on infinite intervals, i.e., $A(t,\lambda) = A_0(t) + \lambda A_1(t)$, $A_0, A_1 \in C([1,\infty])$ and $B(\lambda) \equiv B$. It was shown that all eigenvalues λ of this linear eigenvalue problem, for which the matrix $A(\infty,\lambda) = A_0(\infty) + \lambda A_1(\infty)$ has no eigenvalue on the imaginary axis, are isolated, if not all $\lambda \in \mathbb{C}$ are eigenvalues. Moreover if $A_1(\infty) = 0$ there is an infinite sequence of eigenvalues λ_i with $|\lambda_i| \to \infty$. De Hoog and Weiss (1980a) also proved that the spectral subspaces are finite dimensional.

The first goal of this paper is to show the generalization of the isolatedness statement to nonlinear eigenvalue problems of the form (1.1), (1.2), (1.3). We assume that $B(\lambda)$, $A(t,\lambda)$ are analytic in $\lambda \in \phi \supset \Omega$, where Ω is the domain in which $A(\infty,\lambda)$ has no eigenvalue on the imaginary axis. The analyticity is supposed to hold for all $t \in [1,\infty]$ and $A(t,\lambda)$ is jointly continuous in $[1,\infty] \times \Omega$. The number of rows of the

*Received by the editors July 11, 1981, and in revised form February 15, 1982.

†Institut für Angewandte Mathematik, The Technical University of Vienna, Gusshausstrasse 27, A-1040 Wien, Austria. The work of this author was sponsored by the U. S. Army under contract DAAG29-80-C-0041, and by the Austrian Fund for the Advancement of Research. This material is based upon work supported by the National Science Foundation under grant MCS-7927062.

‡Institut für Angewandte Mathematik, The Technical University of Vienna, Gusshausstrasse 27, A-1040 Wien, Austria.

matrix $B(\lambda)$ is assumed to equal r_-, which is the sum of algebraic multiplicities of all eigenvalues of $A(\infty,\lambda)$ with negative real part for $\lambda\in\Omega$.

The second goal of this paper is to investigate the approximating eigenvalue problems

$$x_T'=t^\alpha A(t,\lambda)x_T, \qquad 1\le t\le T, \quad T\gg 1, \tag{1.4}$$

$$B(\lambda)x_T(1)=0, \tag{1.5}$$

$$S(\lambda)x_T(T)=0 \tag{1.6}$$

where $S(\lambda)$ is a suitably chosen matrix with $r_+=n-r_-$ rows.

The main question arising here is to determine which matrices $S(\lambda)$ lead to convergence of the eigenvalues and eigenfunctions of these approximating problems to the eigenvalues and eigenfunctions of (1.1), (1.2), (1.3) as $T\to\infty$. A class of matrices $S(\lambda)$ which implies exponential convergence will be identified. The convergence results are the generalization of the results obtained by Markowich (1982a) for linear eigenvalue problems. As Markowich pointed out, there is not always (even in the case of a linear eigenvalue problem) an obvious way to choose the suitable S which is independent of λ. However, there is an intrinsic way (see Keller (1976)) to set up an "asymptotic" boundary condition S depending (nonlinearly) on λ. Therefore these "finite" eigenvalue problems are, even in the case of a linear "infinite" problem, nonlinear.

This paper is organized as follows. In §2 nonlinear finite-dimensional eigenvalue problems are discussed; §3 is concerned with the case when A is independent of t; in §4 this restriction is dropped, and §5 contains examples illustrating the theory.

2. Finite dimensional nonlinear eigenvalue problems. Let $A(\lambda)$ be a $k\times k$ matrix, holomorphic in some domain $\Omega\subset\mathbb{C}$. A value $\mu\in\Omega$ for which the linear equation

$$A(\mu)\xi=0, \qquad \xi\neq 0, \tag{2.1}$$

has a solution is called an eigenvalue and ξ is a corresponding eigenvector. Let $\det A(\lambda)$ be the determinant of $A(\lambda)$. Since (2.1) holds if and only if $\det A(\mu)=0$, it follows from the identity theorem of holomorphic functions that either all $\lambda\in\Omega$ are eigenvalues or every compact subset of Ω contains at most finitely many eigenvalues.

Let $\varepsilon\in(0,\varepsilon_0]$ be a real parameter and $B(\lambda,\varepsilon)$ be a $k\times k$ matrix, holomorphic in Ω for all $\varepsilon\in(0,\varepsilon_0]$, with

$$\lim_{\varepsilon\to 0}\sup_{\lambda\in\Lambda}\|B(\lambda,\varepsilon)\|=0 \quad \text{for all } \Lambda \text{ compact}, \quad \Lambda\subset\Omega \tag{2.2}$$

where $\|\cdot\|$ denotes some matrix norm. Now consider the perturbed nonlinear eigenvalue problem

$$C(\lambda,\varepsilon)\xi\equiv(A(\lambda)+B(\lambda,\varepsilon))\xi=0, \qquad \xi\neq 0. \tag{2.3}$$

Since

$$\lim_{\varepsilon\to 0}\det C(\lambda,\varepsilon)=\det A(\lambda), \tag{2.4}$$

we may employ standard perturbation results for zeros of holomorphic functions. Let μ be a root of order s of $\det A(\lambda)=0$, θ be a neighbourhood of μ and

$$b(\varepsilon)=\sup_{\lambda\in\theta}|\det A(\lambda)-\det C(\lambda,\varepsilon)|. \tag{2.5}$$

Then we get

THEOREM 2.1. (i) *When* ε *is sufficiently small there are precisely* s *eigenvalues* $\mu_\varepsilon^1,\cdots,\mu_\varepsilon^s$ *of* (2.3) *near* μ (*counting multiplicities*) *and they satisfy*

$$|\mu_\varepsilon^j-\mu|\leq \text{const.}\, b(\varepsilon)^{1/s}, \qquad j=1,\cdots,s. \tag{2.6}$$

(ii) *The mean*

$$\mu_\varepsilon=\frac{1}{s}\sum_{i=1}^{s}\mu_\varepsilon^i \tag{2.7}$$

satisfies

$$|\mu_\varepsilon-\mu|\leq \text{const.}\, b(\varepsilon). \tag{2.8}$$

The perturbation statement for the eigenvectors is weaker.

THEOREM 2.2. *Let* $\varepsilon_n\to 0$ *as* $n\to\infty$ *and let* ξ_{ε_n} *be a sequence of eigenvectors of* (2.3) (*with norm one*), *each of them belonging to a* $\mu_{\varepsilon_n}^i$ *for* $i=1,\cdots,s$. *Then there is a subsequence* $\xi_{\varepsilon_{n_k}}$ *which converges to an eigenvector* ξ *of* (2.1) *with norm one and*

$$\inf_{\eta\in N(A(\mu))}\|\eta-\xi_{\varepsilon_n}\|\leq \text{const.}\, b(\varepsilon_n)^{1/s}. \tag{2.9}$$

$N(A(\mu))$ *denotes the nullspace of* $A(\mu)$.

A proof can be found in G. Vainikko (1976, Chap. 4).

In the case of a linear eigenvalue problem $A(\lambda)\equiv A-\lambda I$, we get a stronger perturbation result for eigenvectors if the algebraic and geometric multiplicity of the eigenvalue μ is equal to one. Therefore we define:

DEFINITION 2.1. The eigenvalue μ of (2.1) is called simple if μ is a zero of order one of $\det A(\lambda)=0$.

It is easily seen that μ is a simple eigenvalue of (2.1) if and only if $\det(A(\mu)+\tau A'(\mu))=0$ has a zero of order one at $\tau=0$. This again holds if and only if $\tau=0$ is an eigenvalue of geometric and algebraic multiplicity one of the generalized linear eigenvalue problem

$$(A(\mu)+\tau A'(\mu))\xi=0. \tag{2.10}$$

We have

THEOREM 2.3. *Let* μ *be a simple eigenvalue of* (2.1), ξ *a corresponding eigenvector of norm one. Then*:

(i) *for* ε *sufficiently small there is a unique eigenvalue* μ_ε *of* (2.3), *and it satisfies*

$$|\mu_\varepsilon-\mu|\leq \text{const.}|\det C(\mu,\varepsilon)|; \tag{2.11}$$

(ii) *for every* μ_ε *there is exactly one eigenvector* ξ_ε (*with norm one*) *of* (2.3), *and this eigenvector satisfies*

$$\|\xi_\varepsilon-\xi\|\leq \text{const.}|\det C(\mu,\varepsilon)|. \tag{2.12}$$

Proof. The equations

$$C(\lambda,\varepsilon)\xi_\varepsilon=0, \qquad \|\xi_\varepsilon\|=1 \tag{2.13}$$

where $\|\cdot\|$ indicates the Euclidean norm in $\mathbb{C}^k$ are a nonlinear system of equations for $(\mu_\varepsilon,\xi_\varepsilon)$. The Fréchet derivative of the unperturbed problem ($\varepsilon=0$) at (μ,ξ) is given by

the matrix

$$\begin{bmatrix} A(\mu) & A'(\mu)\xi \\ \xi^T & 0 \end{bmatrix}, \tag{2.14}$$

which is nonsingular because μ is simple. (i) and (ii) follow in a straightforward way by applying the techniques of Keller (1975) and Vainikko (1976, Chap. 4). Here θ shrinks to the point μ. □

We conclude this section with a result on holomorphic families of projections.

THEOREM 2.4. *Let $P(\lambda):\mathbb{C}^k\to\mathbb{C}^k$ be a family of projections, holomorphic for $\lambda\in\Omega$. Then*:

(i) *for any pair $(\lambda_1,\lambda_2)\in\Omega\times\Omega$ there is a nonsingular $n\times n$ matrix $Q(\lambda_1,\lambda_2)$ such that*

$$P(\lambda_1)=Q(\lambda_1,\lambda_2)^{-1}P(\lambda_2)Q(\lambda_1,\lambda_2); \tag{2.15}$$

(ii) *$P(\lambda_1)C^k$ is isomorphic to $P(\lambda_2)\mathbb{C}^k$ for all $\lambda_1,\lambda_2\in\Omega$;*

(iii) *Let $r=\operatorname{rank}P(\lambda)$. Then there is a $k\times k$ matrix of rank r, holomorphic in Ω, whose columns span $P(\lambda)\mathbb{C}^k$.*

Proof. (i) follows from Kato (1966) and (ii), (iii) follow easily from (i). □

3. Nonlinear constant-coefficient eigenvalue problems. We consider

$$y'=t^\alpha A(\lambda)y, \qquad 1\le t<\infty, \quad \alpha>-1, \tag{3.1}$$

$$B(\lambda)y(1)=0, \tag{3.2}$$

$$y\in C([1,\infty]) \tag{3.3}$$

where $A(\lambda)$ is an $n\times n$ matrix.

The analysis for these problems will outline the approach for the more complicated case, when A is also a function of the independent variable t. We assume that A,B are holomorphic in some domain ϕ in the complex plane and that there is a domain $\Omega\subset\phi$, so that $A(\lambda)$ has no eigenvalue $\nu(\lambda)$ with vanishing real part when $\lambda\in\Omega$. Then, for all $\lambda\in\Omega$, $A(\lambda)$ has a fixed number of eigenvalues with negative real part, which we call r_-, and a fixed number of eigenvalues with positive real part, which we call r_+ ($r_++r_-=n$). Now we take a compact subset $\Lambda\subset\Omega$. Then there are two closed rectifiable curves Γ_+, Γ_-, completely in the right and left half planes respectively, so that for all $\lambda\in\Lambda$ all eigenvalues of $A(\lambda)$ are enclosed by either Γ_+ or Γ_-.

Now let

$$P_+(\lambda)=\frac{1}{2\pi i}\int_{\Gamma_+}(z-A(\lambda))^{-1}dz, \qquad \operatorname{rank}P_+(\lambda)=r_+, \tag{3.4}$$

$$P_-(\lambda)=\frac{1}{2\pi i}\int_{\Gamma_-}(z-A(\lambda))^{-1}dz, \qquad \operatorname{rank}P_-(\lambda)=r_- \tag{3.5}$$

be the total projections onto the direct sum of invariant subspaces associated with eigenvalues of $A(\lambda)$ with positive and negative real parts, respectively.

From Kato (1966, Chap. 2) we conclude that P_+, P_- are holomorphic in Λ^0, the interior of Λ.

The general solution of the problem (3.1), (3.3) is

$$y(t,\lambda)=\exp\left(\frac{t^{\alpha+1}}{\alpha+1}A(\lambda)\right)P_-(\lambda)\xi, \qquad \xi\in\mathbb{C}^n. \tag{3.6}$$

Theorem 2.4(iii), implies that there is an $n\times r_-$ matrix $V(\lambda)$ of full rank and holomorphic in Λ^0 which spans $P_-(\lambda)\mathbb{C}^n$. Using V in (3.6) and inserting into the boundary condition (3.2), we obtain

$$F(\lambda)\eta\equiv B(\lambda)\exp\left(\frac{A(\lambda)}{\alpha+1}\right)V(\lambda)\eta=0,\qquad \eta\in\mathbb{C}^{r_-} \tag{3.7}$$

assuming that $B(\lambda)$ is an $r_-\times n$ matrix. Every pair (μ,η), $\eta\neq 0$ which solves (3.7) determines a solution of the eigenvalue problem (3.1), (3.2), (3.3) by

$$y(t,\mu)=\exp\left(\frac{t^{\alpha+1}}{\alpha+1}A(\mu)\right)V(\mu)\eta. \tag{3.8}$$

Our assumptions guarantee that $F(\lambda)\equiv B(\lambda)\exp(A(\lambda)/(\alpha+1))V(\lambda)$ is holomorphic in Λ^0, so we obtain from §2:

THEOREM 3.1. *Let $B(\lambda)$ be an $r_-\times n$ matrix holomorphic in ϕ. Then either all $\lambda\in\Omega$ are eigenvalues of* (3.1), (3.2), (3.3) *or every compact subset of Ω contains at most a finite number of eigenvalues. If μ is an eigenvalue of* (3.1), (3.2), (3.3), *the dimension of the nullspace is between* 1 *and* r_-.

Now we approximate the eigenvalue problem (3.1), (3.2), (3.3) by finite interval problems:

$$x_T'=t^\alpha A(\lambda)x_T,\qquad 1\le t\le T,\quad T\gg 1, \tag{3.9}$$

$$B(\lambda)x_T(1)=0, \tag{3.10}$$

$$S(\lambda)x_T(T)=0 \tag{3.11}$$

where $S(\lambda)$ is an $r_+\times n$ matrix. Choices of $S(\lambda)$ will be discussed later.

We write the general solution of (3.9) as

$$x_T(t)=\exp\left(\frac{t^{\alpha+1}}{\alpha+1}A(\lambda)\right)V(\lambda)\eta_-+\exp\left(\frac{t^{\alpha+1}-T^{\alpha+1}}{\alpha+1}A(\lambda)\right)W(\lambda)\eta_+ \tag{3.12}$$

where the columns of the $n\times r_+$ matrix $W(\lambda)$ span $P_+(\lambda)\mathbb{C}^n$ and are holomorphic in Λ^0. The use of (3.12) in (3.10), (3.11) yields the $n\times n$ block system

(3.13)

$$F(\lambda,T)\begin{pmatrix}\eta_-\\ \eta_+\end{pmatrix}=\begin{bmatrix} B(\lambda)\exp\left(\frac{A(\lambda)}{\alpha+1}\right)V(\lambda) & B(\lambda)\exp\left(\frac{1-T^{\alpha+1}}{\alpha+1}A(\lambda)\right)W(\lambda)\\ S(\lambda)\exp\left(\frac{T^{\alpha+1}}{\alpha+1}A(\lambda)\right)V(\lambda) & S(\lambda)W(\lambda)\end{bmatrix}\begin{pmatrix}\eta_-\\ \eta_+\end{pmatrix}=0.$$

By (3.7) we conclude that

$$\det F(\lambda,T)=\det F(\lambda)\cdot\det S(\lambda)W(\lambda)+c(\lambda,T) \tag{3.14}$$

holds, where

$$|c(\lambda,T)|\le\text{const.}\left\|\exp\left(\frac{-T^{\alpha+1}}{\alpha+1}A(\lambda)\right)W(\lambda)\right\|\cdot\left\|S(\lambda)\exp\left(\frac{T^{\alpha+1}}{\alpha+1}A(\lambda)\right)V(\lambda)\right\|. \tag{3.15}$$

Let $\nu_-(\lambda)$ be the largest negative real part of the eigenvalues of $A(\lambda)$ and let $\nu_+(\lambda)$ be the smallest positive real part of the eigenvalues. Then (3.15) reduces to

$$|c(\lambda,T)|\leq \text{const.}(\lambda,\rho)\exp\left((\nu_-(\lambda)-\nu_+(\lambda)+\rho)\frac{T^{\alpha+1}}{\alpha+1}\right) \tag{3.16}$$

where const.(λ,ρ) is bounded when λ varies in a compact set and $\rho>0$. Now we prove the convergence theorem.

THEOREM 3.2. *Let the $r_+\times n$ matrix $S(\lambda)$ be holomorphic for $\lambda\in\Omega$ and assume that* $\det S(\lambda)W(\lambda)\neq 0$ *for* $\lambda\in\Omega$. *Let $\mu\in\Omega$ be an eigenvalue of* (3.1), (3.2), (3.3) *of order s, i.e.,* $\det F(\lambda)$ *has a zero at $\lambda=\mu$ of order s. Then there are exactly s eigenvalues $\mu_T^1,\cdots,\mu_T^s$ (counting multiplicities of the zeros of* $\det F(\lambda,T)$*) for T sufficiently large in a sufficiently small neighbourhood of μ, and for all $\rho>0$ there is a constant depending on ρ such that*

$$\max_{i=1(1)s}|\mu_T^i-\mu|\leq \text{const.}(\rho)\exp\left((\nu_-(\mu)-\nu_+(\mu)+\rho)\frac{T^{\alpha+1}}{s(\alpha+1)}\right), \tag{3.17}$$

$$|\mu_T-\mu|\leq \text{const.}(\rho)\exp\left((\nu_-(\mu)-\nu_+(\mu)+\rho)\frac{T^{\alpha+1}}{\alpha+1}\right) \tag{3.18}$$

where $\mu_T=\frac{1}{s}\sum_{i=1}^s\mu_T^i$. Let x_T be an eigenfunction belonging to one of the μ_T^i's. Then

$$\inf_{y\in N_\mu}\|x_T-y\|_{[1,T]}\leq \text{const.}(\rho)\exp\left((\nu_-(\mu)-\nu_+(\mu)+\rho)\frac{T^{\alpha+1}}{s(\alpha+1)}\right) \tag{3.19}$$

where N_μ denotes the nullspace of (3.1), (3.2), (3.3) *for $\lambda=\mu$.*

We denote $\|f\|_{[a,b]}=\max_{t\in[a,b]}\|f(t)\|$ for $f\in C([a,b])$, $a<b\leq\infty$.

Proof. All statements follow immediately by regarding (3.13) as a perturbation of the eigenvalue problem

$$\begin{bmatrix} B(\lambda)\exp\left(\frac{A(\lambda)}{\alpha+1}\right)V(\lambda) & 0 \\ 0 & S(\lambda)W(\lambda)\end{bmatrix}\begin{pmatrix}\eta\\ \xi\end{pmatrix}=0 \tag{3.20}$$

(which is equivalent to (3.7)) and by applying the Theorems 2.1 and 2.2.

Now we discuss a possible choice of $S(\lambda)$. Let the rows of the $r_+\times n$ matrix $S_p(\lambda)$ span the range of $(P_+(\lambda))^T$ (the superscript T denotes transposition). Then the asymptotic boundary condition

$$S_p(\lambda)x_T(T)=0 \tag{3.21}$$

fulfills the assumptions of Theorem 3.2. Moreover,

$$P_+(\lambda)\exp\left(\frac{T^{\alpha+1}}{\alpha+1}A(\lambda)\right)V(\lambda)\equiv 0,\qquad \lambda\in\Omega, \tag{3.22}$$

holds. Therefore, when using the boundary condition (3.21), the (2.1) portion of the matrix in (3.13) vanishes for all T and the approximate problems (3.9), (3.10), (3.11) reproduce the eigenvalues and eigenfunctions of the problem (3.1), (3.2), (3.3) exactly. □

4. General nonlinear eigenvalue problems on infinite intervals. We consider the problem

$$y'=t^\alpha A(t,\lambda)y,\qquad 1\leq t<\infty,\quad \alpha>-1, \tag{4.1}$$

$$B(\lambda)y(1)=0, \tag{4.2}$$

$$y\in C([1,\infty]) \tag{4.3}$$

where $A(t,\lambda)$ is an $n\times n$ matrix holomorphic for λ in some domain ϕ and every fixed $t\in[1,\infty]$ and continuous in $[1,\infty]\times\phi$. Also B is holomorphic in ϕ. We assume that there is a domain $\Omega\subset\phi$, so that the matrix $A(\infty,\lambda)$ has no eigenvalue $\nu(\lambda)$ on the imaginary axis for $\lambda\in\Omega$. As in §3 we take any compact subset $\Lambda\subset\Omega$ and construct the projections $P_+(\lambda)$, $P_-(\lambda)$

$$P_+(\lambda)=\frac{1}{2\pi i}\int_{\Gamma_+}(z-A(\infty,\lambda))^{-1}dz, \qquad \operatorname{rank}P_+(\lambda)\equiv r_+, \tag{4.4}$$

$$P_-(\lambda)=\frac{1}{2\pi i}\int_{\Gamma_-}(z-A(\infty,\lambda))^{-1}dz, \qquad \operatorname{rank}P_-(\lambda)\equiv r_-. \tag{4.5}$$

The contours Γ_+, Γ_- are chosen as in §3. We set

$$\phi(t,\lambda)=\exp\left(\frac{A(\infty,\lambda)}{\alpha+1}t^{\alpha+1}\right) \tag{4.6}$$

and define the operator $H_\lambda: C([\delta,\infty])\to C([\delta,\infty])$ for $\delta\geq 1$ by

$$\begin{aligned}(H_\lambda g)(t)=&\phi(t,\lambda)\int_\infty^t P_+(\lambda)\phi^{-1}(s,\lambda)s^\alpha g(s)\,ds\\ &+\phi(t,\lambda)\int_\delta^t P_-(\lambda)\phi^{-1}(s,\lambda)s^\alpha g(s)\,ds\end{aligned} \tag{4.7}$$

so that $H_\lambda g\in C([\delta,\infty])$ is a particular solution of the problem

$$y'=t^\alpha A(\infty,\lambda)y+t^\alpha g(t), \qquad t\geq\delta, \quad g\in C([\delta,\infty]). \tag{4.8}$$

An analysis of H_λ can be found in de Hoog and Weiss (1980a, b). P_+, P_- are holomorphic in Λ^0 for every (fixed) $t\in[\delta,\infty]$ and continuous for $t\in[1,\infty]$. Then it is easy to show that $(H_\lambda g(\cdot,\lambda))(t)$ is holomorphic in Λ^0 for every fixed $t\in[\delta,\infty]$. From de Hoog and Weiss (1980a, b) we conclude that

$$\|H_\lambda\|_{[\delta,\infty]}\leq C(\lambda) \tag{4.9}$$

where $C(\lambda)$ is independent of δ (for $G: C([a,b])\to C([a,b])$ we denote by $\|G\|_{[a,b]}$ the operator norm induced by $\|\cdot\|_{[a,b]}$).

Now we show that $C(\lambda)$ remains bounded when λ varies in compact subsets $K\subset\Lambda^0$. From (4.7) we derive

$$\|H_\lambda\|_{[\delta,\infty]}\leq\max_{t\in[\delta,\infty]}\left(\int_t^\infty\|F_+(t,s,\lambda)\|\,ds+\int_\delta^t\|F_-(t,s,\lambda)\|\,ds\right), \tag{4.10}$$

where

$$F_+(t,s,\lambda)=s^\alpha\phi(t,\lambda)P_+(\lambda)\phi^{-1}(s,\lambda), \qquad \lambda\in\Lambda, \tag{4.11a}$$

$$F_-(t,s,\lambda)=s^\alpha\phi(t,\lambda)P_-(\lambda)\phi^{-1}(s,\lambda), \qquad \lambda\in\Lambda, \tag{4.11b}$$

hold. Obviously, F_+, F_- are holomorphic in Λ^0 for fixed s,t. We transform $A(\infty,\lambda)$ to its Jordan canonical form $J(\infty,\lambda)$:

$$A(\infty,\lambda)=E(\lambda)J(\infty,\lambda)E^{-1}(\lambda), \qquad \lambda\in\Omega, \tag{4.12}$$

and assume that $J(\infty,\lambda)$ has the block structure

$$J(\infty,\lambda)=\operatorname{diag}(J_\infty^+(\lambda),J_\infty^-(\lambda)) \tag{4.13}$$

where the $r_+ \times r_+$ matrix $J_\infty^+(\lambda)$ contains all eigenvalues with positive real part and the $r_- \times r_-$ matrix $J_\infty^-(\lambda)$ contains the eigenvalues with negative real part for all $\lambda \in \Omega$. Defining the diagonal projections

$$D_+ = \operatorname{diag}(I_{r_+}, 0), \qquad D_- = \operatorname{diag}(0, I_{r_-}), \tag{4.14}$$

we have

$$P_+(\lambda) = E(\lambda) D_+ E^{-1}(\lambda), \quad P_-(\lambda) = E(\lambda) D_- E^{-1}(\lambda), \qquad \lambda \in \Omega, \tag{4.15}$$

and hence

$$F_+(t,s,\lambda) = s^\alpha E(\lambda) \begin{bmatrix} \exp\left(\frac{J_\infty^+(\lambda)}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) & 0 \\ 0 & 0 \end{bmatrix} E^{-1}(\lambda) \tag{4.16a}$$

$$F_-(t,s,\lambda) = s^\alpha E(\lambda) \begin{bmatrix} 0 & 0 \\ 0 & \exp\left(\frac{J_\infty^-(\lambda)}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) \end{bmatrix} E^{-1}(\lambda) \tag{4.16b}$$

Each entry of F_+, F_- is a sum of the form

$$f_\pm(t,s,\lambda) = s^\alpha \sum_{i=1}^{r_\pm} a_i^\pm(\lambda) \exp\left(\frac{\nu_i^\pm(\lambda)}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) (t^{\alpha+1} - s^{\alpha+1})^{j_i} \tag{4.17}$$

where $\nu_i^+(\lambda)$ are the eigenvalues of $J_\infty^+(\lambda)$, and $\nu_i^-(\lambda)$ are the eigenvalues of $J_\infty^-(\lambda)$. The integers j_i satisfy $0 \le j_i \le (r_+ - 1)$ and $0 \le j_i \le (r_- - 1)$ respectively. $a_i^\pm(\lambda)$ is a sum of products of elements of $E(\lambda)$ and $E^{-1}(\lambda)$.

Now we take a compact subset $K \subset \Lambda^0$. It follows from Kato (1966) that $E(\lambda)$, $E^{-1}(\lambda)$ can be chosen boundedly in $\tilde{K} = K - \bigcup_{i=1}^N \cup(z_i)$ where the z_i are points at which eigenvalues $\nu_i^\pm(\lambda)$ change algebraic or geometric multiplicities and $\cup(z_i)$ stands for a sufficiently small neighbourhood of z_i. Also $J(\lambda)$ is bounded in $\tilde{K}$. Without loss of generality we assume that none of the z_i's lies on the boundary ∂K, since, if that happens, we can choose a larger compact set $\hat{K} \supset K$, so that $\{z_1, \cdots, z_N\} \cap \partial \hat{K} = \varnothing$. By (4.11) the entries $f_\pm(t,s,\lambda)$ are holomorphic in K; therefore, they take their maximum at the boundary ∂K. The coefficients $a_i(\lambda)$ and the $\nu_i^\pm(\lambda)$ are bounded on ∂K and therefore

$$\max_{\lambda \in K} |f_\pm(t,s,\lambda)| \le r_\pm \cdot s^\alpha \max_{\substack{\lambda \in \partial K \\ j = 1(1) r_\pm}} |a_j^\pm(\lambda)| \max_{0 \le i \le r_\pm} \left(\exp\left(\frac{c_i^\pm}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) \cdot |t^{\alpha+1} - s^{\alpha+1}|^i \right) \tag{4.18}$$

where $c_i^+ = \min_{\lambda \in \partial K} \operatorname{Re} \nu_i^+(\lambda)$, $c_i^- = \max_{\lambda \in \partial K} \operatorname{Re} \nu_i^-(\lambda)$ hold. Therefore we get

$$\max_{\lambda \in K} \|F_+(t,s,\lambda)\| \le c_1(K) s^\alpha \max_{0 \le i \le r_+} \left(\exp\left(\frac{c_+}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) (s^{\alpha+1} - t^{\alpha+1})^i \right) \tag{4.19a}$$

and

$$\max_{\lambda \in K} \|F_-(t,s,\lambda)\| \le c_2(K) s^\alpha \max_{0 \le i \le r_-} \left(\exp\left(\frac{c_-}{\alpha+1} (t^{\alpha+1} - s^{\alpha+1}) \right) (t^{\alpha+1} - s^{\alpha+1})^i \right) \tag{4.19b}$$

where $0 < c_+ = \min_{i = 1(1) r_+} c_i^+$, $0 > c_- = \max_{i = 1(1) r_-} c_i^-$ and $0 \le i \le n$.

Using (4.10) and the estimates derived in Markowich (1983, §1), we get

$$\max_{\lambda\in K}\|H_\lambda\|_{[\delta,\infty]}\le C(K) \tag{4.20}$$

where $C(K)$ is independent of δ.

We rewrite (4.1) as

$$y'=t^\alpha A(\infty,\lambda)y+t^\alpha(A(t,\lambda)-A(\infty,\lambda))y. \tag{4.21}$$

Setting

$$G(t,\lambda)=A(t,\lambda)-A(\infty,\lambda), \tag{4.22}$$

we get from (4.21)

$$y(t)=\phi(t,\lambda)V(\lambda)\eta+(H_\lambda G(\cdot,\lambda)y)(t),\qquad \eta\in\mathbb{C}^{r_-}, \tag{4.23}$$

where $V(\lambda)$ is as in §2. The assumptions on $A(t,\lambda)$ guarantee that there is a $\delta\ge 1$, $\delta=\delta(K)$, such that

$$\|G(\cdot,\lambda)\|_{[\delta,\infty]}\le\frac{1}{2C(K)}\quad\text{for all }\lambda\in K \tag{4.24}$$

where K is any compact subset of Λ^0 and $C(K)$ is the constant defined in (4.20). Then

$$\max_{\lambda\in K}\|H_\lambda G(\cdot,\lambda)\|_{[\delta,\infty]}\le\frac{1}{2}. \tag{4.25}$$

This implies that $I-H_\lambda G(\cdot,\lambda): C([\delta,\infty])\to C([\delta,\infty])$ is nonsingular for all $\lambda\in K$ and

$$y=(I-H_\lambda G(\cdot,\lambda))^{-1}\phi(\cdot,\lambda)V(\lambda)\eta,\qquad \eta\in\mathbb{C}^{r_-} \tag{4.26}$$

holds. y is defined for $t\in[\delta,\infty]$ and all $\lambda\in K$. The series expansion of the $n\times r_-$ matrix

$$\psi_-(t,\lambda)=((I-H_\lambda G(\cdot,\lambda))V(\lambda))(t) \tag{4.27}$$

is given by

$$\psi_-(\cdot,\lambda)=\sum_{i=0}^{\infty}(H_\lambda G(\cdot,\lambda))^i\phi(\cdot,\lambda)V(\lambda)\in C([\delta,\infty]),\qquad \lambda\in K. \tag{4.28}$$

The partial sums $\psi_-^{(k)}(t,\lambda)$ of this series are holomorphic in $\lambda\in\mathring{K}$ for all fixed $t\in[\delta,\infty]$. Because of (4.25) we get

$$\|\psi_-^{(k)}(t,\lambda)\|\le\sum_{i=0}^{k}\left(\frac{1}{2}\right)^i\max_{\lambda\in K}\|\phi(\cdot,\lambda)V(\lambda)\|_{[\delta,\infty]}. \tag{4.29}$$

Since $\phi(t,\lambda)V(\lambda)$ is continuous in both variables, the partial sums are uniformly bounded on K and so $\psi_-(t,\lambda)$ is holomorphic in λ for $\lambda\in\mathring{K}$, and for all fixed $t\in[\delta,\infty]$. By continuation $\psi_-(t,\lambda)$ is holomorphic in $\lambda\in\mathring{K}$ for all fixed $t\in[1,\infty]$.

Inserting (4.26) into the boundary condition (4.2) we obtain the finite dimensional eigenvalue problem

$$F(\lambda)\eta\equiv B(\lambda)\psi_-(1,\lambda)\eta=0,\qquad \eta\in\mathbb{C}^{r_-} \tag{4.30}$$

where $B(\lambda)$ is assumed to be an $r_-\times n$ matrix. Given now any compact subset $\theta\subset\Omega$ we choose Λ, K such that $\Lambda^0\supset K$, $\mathring{K}\supset\theta$. So $F(\lambda)$ is holomorphic in θ and Theorem 3.1 holds for the problem (4.1), (4.2), (4.3). Therefore, excluding the trivial case, all eigenvalues in Ω are isolated and the dimension of the nullspace is between 1 and r_-.

Now we prove the asymptotic estimate for $\psi_-(t,\lambda)$:

$$\max_{\lambda\in\theta}\|\psi_-(t,\lambda)\|\leq \text{const.}\exp\left(\left(\max_{\substack{\lambda\in\partial\theta\\ i=1(1)r_-}}\operatorname{Re}\nu_i^-(\lambda)+\rho\right)\frac{t^{\alpha+1}}{\alpha+1}\right) \tag{4.31}$$

where $\nu_i^-(\lambda)$ are the eigenvalues of $A(\infty,\lambda)$ with real part less than zero, and $\rho>0$ is arbitrary but sufficiently small so that the exponent has negative sign.

If there is one (or more) of the singularities $\{z_1,\cdots,z_N\}$ of $E(\lambda)$ on the boundary $\partial\theta$, we take a larger set θ_1 such that $\theta\subset\theta_1$, $\theta_1\subset K$ and $\{z_1,\cdots,z_N\}\cap\partial\theta_1=\varnothing$. Then we derive as in (4.18):

$$\max_{\lambda\in\theta}\|\phi(t,\lambda)V(\lambda)\|\leq \text{const.}\exp\left(\left(\max_{\substack{\lambda\in\partial\theta\\ i=1(1)r_-}}\operatorname{Re}\nu_i^-(\lambda)+\rho\right)\frac{t^{\alpha+1}}{\alpha+1}\right). \tag{4.32}$$

The necessity to add $\rho>0$ in the exponent comes from the possibility that θ might have to be changed to θ_1 as described above and from the possible occurrence of powers of $|t^{\alpha+1}-s^{\alpha+1}|$. A sufficiently small change and the continuity of the eigenvalues assure that ρ is arbitrarily small.

Using (4.7), (4.11) we have

(4.33)

$$\begin{aligned}\max_{\lambda\in\theta}&\|(H_\lambda G(\cdot,\lambda)\phi(\cdot,\lambda)V(\lambda))(t)\|\\ &\leq\max_{\lambda\in\theta}\|G(\cdot,\lambda)\|_{[\delta,\infty]}\left(\int_t^\infty\max_{\lambda\in\theta}\|F_+(t,s,\lambda)\|\max_{\lambda\in\theta}\|\phi(s,\lambda)V(\lambda)\|\,ds\right.\\ &\qquad\left.+\int_\delta^t\max_{\lambda\in\theta}\|F_-(t,s,\lambda)\|\max_{\lambda\in\theta}\|\phi(s,\lambda)V(\lambda)\|\,ds\right).\end{aligned}$$

Now (4.19), (4.32) can be used to bound the right-hand side of (4.33). θ has to be substituted for K in the definition of c_i^+, c_i^-. Since $\rho>0$, the estimate given in Markowich (1983, §1, Thm. 2.3) can be used, and

$$\begin{aligned}&\max_{\lambda\in\theta}\|(H_\lambda G(\cdot,\lambda)\phi(\cdot,\lambda)(t))\|\\ &\qquad\leq\text{const.}\max_{\lambda\in\theta}\|G(\cdot,\lambda)\|_{[\delta,\infty]}\exp\left(\left(\max_{\substack{\lambda\in\partial\theta\\ i=1(1)r_-}}\operatorname{Re}\nu_i^-(\lambda)+\rho\right)\frac{t^{\alpha+1}}{\alpha+1}\right)\end{aligned} \tag{4.34}$$

follows. Repeated use of (4.34) and (4.28) gives (4.31)

As in §3, we investigate the approximating finite eigenvalue problems

$$x_T'=t^\alpha A(t,\lambda)x_T,\qquad 1\leq t\leq T,\quad T\gg 1, \tag{4.35}$$

$$B(\lambda)x_T(1)=0, \tag{4.36}$$

$$S(\lambda)x_T(T)=0 \tag{4.37}$$

where $S(\lambda)$ is a suitably chosen $r_+\times n$ matrix whose entries are holomorphic in Ω. Rewriting (4.35) as

$$x_T'=t^\alpha A(\infty,\lambda)x_T+t^\alpha G(t,\lambda)x_T,\qquad 1\leq t\leq T, \tag{4.38}$$

where $G(t,\lambda)$ is defined as in (4.22), we set

$$x_T=\phi(t,\lambda)V(\lambda)\eta_-+\phi(t,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\eta_++(H_{\lambda,T}G(\cdot,\lambda)x_T)(t) \tag{4.39}$$

where the columns of the $n\times r_+$ matrix $W(\lambda)$, which can be chosen holomorphic in Λ^0, span the range of $P_+(\lambda)$, and $H_{\lambda,T}: C([\delta,T])\to C([\delta,T])$ is defined as

$$H_{\lambda,T}g=H_\lambda g_T \tag{4.40}$$

for $g\in C([\delta,T])$, $1\le\delta\le T$, where

$$g_T(t)=\begin{cases} g(t), & \delta\le t\le T,\\ g(T), & t\ge T,\end{cases} \tag{4.41}$$

has been set.

Given a fixed eigenvalue $\lambda=\mu\in\Omega$ of (4.1), (4.2), (4.3), we take a compact subset $K\subset\Lambda^0$ with $\mu\in K$ and conclude from (4.20)

$$\max_{\lambda\in K}\|H_{\lambda,T}\|_{[\delta,T]}\le\max_{\lambda\in K}\|H_\lambda\|_{[\delta,\infty]}\le C(K). \tag{4.42}$$

Therefore there is a fixed $\delta=\delta(K)\ge 1$ such that

$$\max_{\lambda\in K}\|H_{\lambda,T}G(\cdot,\lambda)\|_{[\delta,T]}\le\frac{1}{2}, \tag{4.43}$$

and so $(I-H_{\lambda,T}G(\cdot,\lambda))^{-1}$ exists for all $\lambda\in K$ as an operator on $C([\delta,T])$. We get from (4.39):

$$\begin{aligned} x_T=&\big(I-H_{\lambda,T}G(\cdot,\lambda)\big)^{-1}\phi(\cdot,\lambda)V(\lambda)\eta_-\\ &+\big(I-H_{\lambda,T}G(\cdot,\lambda)\big)^{-1}\phi(\cdot,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\eta_+\end{aligned} \tag{4.44}$$

on $[\delta,\infty]$. The analyticity of

$$ {}_T\psi_-(t,\lambda)=\Big(\big(I-H_{\lambda,T}G(\cdot,\lambda)\big)^{-1}\phi(\cdot,\lambda)V(\lambda)\Big)(t), \tag{4.45a}$$

$$ {}_T\psi_+(t,\lambda)=\Big(\big(I-H_{\lambda,T}G(\cdot,\lambda)\big)^{-1}\phi(\cdot,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\Big)(t) \tag{4.45b}$$

in λ for $t\in[\delta,T]$ follows as the analyticity of $\psi_-(t,\lambda)$.

The $n\times r_-$ matrix ${}_T\psi_-$ and the $n\times r_+$ matrix ${}_T\psi_+$ respectively satisfy the equations

$$ {}_T\psi_-(t,\lambda)-\big(H_{\lambda,T}G(\cdot,\lambda){}_T\psi_-(\cdot,\lambda)\big)(t)=\phi(t,\lambda)V(\lambda), \tag{4.46a}$$

$$ {}_T\psi_+(t,\lambda)-\big(H_{\lambda,T}G(\cdot,\lambda){}_T\psi_+(\cdot,\lambda)\big)(t)=\phi(t,\lambda)\phi^{-1}(T,\lambda)W(\lambda). \tag{4.46b}$$

Similarly to de Hoog and Weiss (1980a) we derive some properties of ${}_T\psi_-, {}_T\psi_+$. From (4.27) and (4.46a) we get

$$ {}_T\psi_--\psi_-=H_{\lambda,T}G(\cdot,\lambda)({}_T\psi_--\psi_-)+(H_{\lambda,T}G(\cdot,\lambda)\psi_--H_\lambda G(\cdot,\lambda)\psi_-), \tag{4.47}$$

and therefore we get by regarding $G(\cdot,\lambda)\psi_-(\cdot,\lambda)\in C([\delta,T])$

$$ {}_T\psi_-(\cdot,\lambda)-\psi_-(\cdot,\lambda)=\big(I-H_{\lambda,T}G(\cdot,\lambda)\big)^{-1}(H_{\lambda,T}-H_\lambda)G(\cdot,\lambda)\psi_-(\cdot,\lambda)\in C([\delta,T]). \tag{4.48}$$

Obviously for $g\in C([\delta,T])$ and $t\in[\delta,T]$

$$\begin{aligned}\big((H_{\lambda,T}-H_\lambda)g\big)(t)&=\phi(t,\lambda)\int_\infty^T P_+(\lambda)\phi^{-1}(s,\lambda)s^\alpha(g(T)-g(s))\,ds\\ &=\phi(t,\lambda)\phi^{-1}(T,\lambda)P_+(\lambda)\phi(T,\lambda)\\ &\qquad\times\int_\infty^T P_+(\lambda)\phi^{-1}(s,\lambda)s^\alpha(g(T)-g(s))\,ds\\ &=\phi(t,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\gamma(g,T)\end{aligned} \tag{4.49}$$

with $\gamma(g,T)\in C^{r_+}$. So

$$((H_{\lambda,T}-H_\lambda)G(\cdot,\lambda)\psi_-(\cdot,\lambda))(t)=\phi(t,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\Gamma_T$$

where Γ_T is an $r_+\times r_-$ matrix. From (4.46b) and (4.48) we derive

$$(4.50)\qquad {}_T\psi_-(\cdot,\lambda)=\psi_-(\cdot,\lambda)+{}_T\psi_+(\cdot,\lambda)\Gamma_T.$$

Therefore the matrix $[\psi_-(t,\lambda),{}_T\psi_+(t,\lambda)]$ has rank n for all $t\in[\delta,T]$ and is a fundamental matrix of (4.35).

Instead of using (4.44), we can write the general solution of (4.35) as

$$(4.51)\qquad x_T=\psi_-(t,\lambda)\eta_-+{}_T\psi_+(t,\lambda)\eta_+.$$

For the following we need an estimate for ${}_T\psi_+(\cdot,\lambda)$. From (4.45b) we obtain

$$(4.52)\qquad \begin{aligned}&\max_{\lambda\in\theta}\|{}_T\psi_+(\cdot,\lambda)-\phi(\cdot,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\|_{[\delta,T]}\\ &\quad\le\text{const.}\max_{\lambda\in\theta}\|G(\cdot,\lambda)\phi(\cdot,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\|_{[\delta,T]}.\end{aligned}$$

Using similar analyticity arguments as above it is easy to check that the right-hand side of (4.52) can be estimated by

$$(4.53)\qquad w(T,\kappa)=\text{const.}\max_{t\in[\delta,T]}\left(\max_{\lambda\in\theta}\|G(t,\lambda)\|\exp\left(\frac{\kappa-\rho}{\alpha+1}(t^{\alpha+1}-T^{\alpha+1})\right)\right)$$

where $\kappa\equiv\kappa(\theta)=\min_{\lambda\in\partial\theta,i=1(1)r_+}\operatorname{Re}\mu_i^+(\lambda)$ and $\rho>0$ is arbitrarily small.

Obviously $\lim_{T\to\infty}w(T,\kappa)=0$ and we get after continuation to $[1,T]$

$$(4.54)\qquad \lim_{T\to\infty}\|{}_T\psi_+(\cdot,\lambda)-\phi(\cdot,\lambda)\phi^{-1}(T,\lambda)W(\lambda)\|_{[1,T]}=0$$

uniformly for $\lambda\in\theta$.

Now we evaluate (4.51) at the boundaries $t=1,T$, substitute into (4.36), (4.37) and obtain the n-dimensional nonlinear eigenvalue problem

$$(4.55)\qquad F(\lambda,T)\begin{pmatrix}\eta_-\\ \eta_+\end{pmatrix}=\begin{bmatrix}B(\lambda)\psi_-(1,\lambda) & B(\lambda){}_T\psi_+(1,\lambda)\\ S(\lambda)\psi_-(T,\lambda) & S(\lambda){}_T\psi_+(T,\lambda)\end{bmatrix}\begin{pmatrix}\eta_-\\ \eta_+\end{pmatrix}=0.$$

Interpreting (4.55) as a perturbation of

$$(4.56)\qquad \begin{bmatrix}B(\lambda)\psi_-(1,\lambda) & 0\\ 0 & S(\lambda)W(\lambda)+o(T,\lambda)\end{bmatrix}\begin{pmatrix}\eta\\ \xi\end{pmatrix}=0,$$

we get with $F(\lambda)\equiv B(\lambda)\psi_-(1,\lambda)$

$$(4.57)\qquad \begin{aligned}\det F(\lambda,T)=\det F(\lambda)(\det S(\lambda)W(\lambda)+o(T,\lambda))\\ +O(\|{}_T\psi_+(1,\lambda)\|\cdot\|S(\lambda)\psi_-(T,\lambda)\|)\end{aligned}$$

where

$$(4.58)\qquad |o(T,\lambda)|\to 0\quad\text{as } T\to\infty$$

uniformly for $\lambda\in\theta$.

Assuming that $S(\lambda)W(\lambda)$ is nonsingular, we obtain by dividing through $\det S(\lambda)W(\lambda)+o(T,\lambda)$ and by applying the perturbation arguments of §2:

THEOREM 4.1. *Let the $r_+\times n$ matrix $S(\lambda)$ be holomorphic for $\lambda\in\Omega$ and assume that $\det S(\lambda)W(\lambda)\not\equiv 0$ in Ω. Let $\mu\in\Omega$ be an eigenvalue of* (4.1), (4.2), (4.3) *of order s, i.e., $\det F(\lambda)=\det B(\lambda)\psi_-(1,\lambda)$ has a zero order s at $\lambda=\mu$. Then there are exactly s eigenvalues $\mu_T^1,\cdots,\mu_T^s$ (counting multiplicities of the zeros of $\det F(\lambda,T)$) for T sufficiently*

large in a sufficiently small neighbourhood $\bar{S}_\mu$ *of* μ, *and*

$$\max_{i=1(1)s} |\mu_T^i - \mu| < \text{const.}(\rho)\left(w\left(T,\kappa\left(\bar{S}_\mu\right)\right)\right)^{1/s} \exp\left(\left(\nu_-(\mu)+\rho\right)\frac{T^{\alpha+1}}{s(\alpha+1)}\right) \tag{4.59}$$

where $\nu_-(\mu)=\max_{i=1(1)r_-} \operatorname{Re}\nu_i^-(\mu)$ *and* $w(T,\kappa)$ *is defined in* (4.53) *and* $\rho>0$ *is arbitrarily small. Also*

$$|\hat{\mu}_T - \mu| \le \text{const.}(\rho) w\left(T,\kappa\left(\bar{S}_\mu\right)\right) \exp\left(\left(\nu_-(\mu)+\rho\right)\frac{T^{\alpha+1}}{\alpha+1}\right) \tag{4.60}$$

holds where $\hat{\mu}_T=\frac{1}{s}\sum_{i=1}^s \mu_T^i$. *Let* x_T *be an eigenfunction belonging to one of the* μ_T^i*'s. Then*

$$\inf_{y\in N_\mu} \|x_T - y\|_{[1,T]} \le \text{const.}(\rho)\left(w\left(T,\kappa\left(\bar{S}_\mu\right)\right)\right)^{1/s} \exp\left(\left(\nu_-(\mu)+\rho\right)\frac{T^{\alpha+1}}{s(\alpha+1)}\right) \tag{4.61}$$

holds where N_μ *denotes the nullspace of* (4.1), (4.2), (4.3) *for* $\lambda=\mu$.

These convergence results are the extension of the convergence results for linear eigenvalue problems given in Markowich (1982a). The orders of convergence obtained there hold without any change for nonlinear problems.

A possible choice for $S(\lambda)$ is given by (3.21), i.e., the rows of the holomorphic $r_+ \times n$ matrix $S(\lambda)=S_p(\lambda)$ span the range of $(P_+(\lambda))^T$ (the superscript T denotes transposition). This choice reproduces eigenvalues and eigenvectors exactly in the case that A does not depend on t. However, in the general case this does not hold anymore, although in some important cases the asymptotic boundary condition $S_p(\lambda)x_T(T)=0$ implies a faster order of convergence than given in Theorem 4.1. Assume that $A(t,\lambda)$ decays algebraically or exponentially:

$$A(t,\lambda)=A(\infty,\lambda)+O(t^\gamma e^{-a(t)}) \quad \text{for } t\to\infty \tag{4.62}$$

uniformly in compact subset $K\subset\Omega$ where $\gamma\in R$ and $a(t)\ge 0$ is a real function such that $t^\gamma e^{-a(t)}\to 0$ as $t\to\infty$. Then since $S_p(\lambda)\phi(T,\lambda)V(\lambda)\equiv 0$, we get from (4.46a)

$$\begin{aligned}\left\|S_p(\lambda)\psi_-(T,\lambda)\right\| &= \left\|S_p(\lambda)\left(H_\lambda G(\cdot,\lambda)\psi_-(\cdot,\lambda)\right)(T)\right\| \\ &\le \text{const.}\, T^\gamma e^{-a(T)} \exp\left(\left(\max_{i=1(1)r_-} \operatorname{Re}\nu_i^-(\lambda)+\rho\right)\frac{T^{\alpha+1}}{\alpha+1}\right).\end{aligned}$$

This follows from the estimates given in Markowich (1983) applied to (4.27). In this case the right-hand side of the estimate (4.59), (4.61) given in Theorem 4.1 can be multiplied by $(T^\gamma e^{-a(T)})^{1/s}$, and the right-hand side of (4.60) can be multiplied by $T^\gamma e^{-a(T)}$.

Now we consider the case of simple eigenvalues of (4.1), (4.2), (4.3). Since we only defined simple eigenvalues for finite dimensional nonlinear eigenvalue problems, we give

DEFINITION 4.1. An eigenvalue $\mu\in\Omega$ of (4.1), (4.2), (4.3) is called simple if the corresponding nullspace is one dimensional, say it is spanned by the normed vector y, and if the problem

$$v' - t^\alpha A(t,\mu)v = t^\alpha A_\lambda(t,\mu)y(t), \tag{4.63}$$

$$B(\mu)v(1)+B_\lambda(\mu)y(1)=0, \tag{4.64}$$

$$v\in C([1,\infty]) \tag{4.65}$$

has no solution.

Now we show

THEOREM 4.2. *The eigenvalue μ of* (4.1), (4.2), (4.3) *is simple if and only if μ is a first order zero of* $\det F(\lambda)=0$.

Proof. Since y is an eigenvector corresponding to the eigenvalue $\mu\in\Omega$,

$$y(t,\mu)=\psi_-(t,\mu)\xi$$

holds for some $\xi\in\mathbb{C}^{r-}$. Obviously

$$y_\lambda(t,\mu)=\frac{d}{d\lambda}\psi_-(t,\mu)\xi$$

is a particular solution of (4.63), (4.65). So the general solution of (4.63) is

$$v(t)=\psi_-(t,\mu)\beta+y_\lambda(t,\mu),\qquad \beta\in\mathbb{C}^{r-}.$$

Inserting into (4.64) gives

$$B(\mu)\psi_-(1,\mu)\beta=-\left(B(\mu)\frac{d}{d\lambda}\psi_-(1,\mu)+B_\lambda(\mu)\psi_-(1,\mu)\right)\xi$$

or

$$F(\mu)\beta=-F_\lambda(\mu)\xi.$$

This equation is unsolvable (for β) if and only if the generalized linear eigenvalue problem $(F(\mu)+\kappa F_\lambda(\mu))\eta=0$ has $\kappa=0$ as an eigenvalue with geometric and algebraic multiplicity 1. This again holds if and only if $\det F(\lambda)=0$ has a first order zero at $\lambda=\mu$. □

Now we show that approximations for an eigenvalue-eigenvector pair (λ,y) can be computed—in the case where λ is simple—as solutions to nonlinear "finite" two-point boundary value problems. Lentini and Keller (1980) did computations pursuing this way.

We set, assuming that $\mu\in\Omega$ is simple,

$$y_{n+1}=\mu,\quad z=(y_1,\cdots,y_n,y_{n+1})^T,\quad \sum_{i=1}^{n}y_i^2(1)=1 \tag{4.66}$$

(the superscript T denotes transposition) and get from (4.1), (4.2), (4.3)

$$z'=t^\alpha\begin{bmatrix}A(t,z_{n+1})\begin{pmatrix}z_1\\ \vdots\\ z_n\end{pmatrix}\\ 0\end{bmatrix}\equiv t^\alpha f(t,z),\qquad 1\le t<\infty, \tag{4.67}$$

$$\begin{bmatrix}B(z_{n+1}(1))\begin{pmatrix}z_1(1)\\ \vdots\\ z_n(1)\end{pmatrix}\\ \sum_{i=1}^{n}z_i^2(1)-1\end{bmatrix}\equiv b(z(1))=0, \tag{4.68}$$

$$z\in C([1,\infty]). \tag{4.69}$$

The condition $\Sigma_{i=1}^n y_i^2(1)=1$ shall sort one eigenfunction $y\not\equiv 0$ out of the one dimensional eigenspace.

Equations (4.67) and (4.68), (4.69) form a singular two-point boundary value problem as described by de Hoog and Weiss (1980a, b), Markowich (1982a, b), (1983) and Lentini and Keller (1980). Since all eigenvalues in Ω are isolated and since μ is simple, the solution $z=(y_1,\cdots,y_n,\mu)^T$ of (4.67), (4.68), (4.69) is locally unique. Now we will show that z is isolated, i.e., the linearized problem has only the zero solution. We get for the linearized problem with $u=(u_1,\cdots,u_n,u_{n+1})^T$

$$u'=t^{\alpha}\begin{bmatrix} A(t,\mu) & A_{\lambda}(t,\mu)y(t) \\ 0 & 0 \end{bmatrix}u, \tag{4.70}$$

$$\begin{bmatrix} B(\mu) & B_{\lambda}(\mu)y(1) \\ 2y(1)^T & 0 \end{bmatrix}u(1)=0, \tag{4.71}$$

$$u\in C([1,\infty]). \tag{4.72}$$

Setting $v=(u_1,\cdots,u_n)^T$ we derive

$$v'-t^{\alpha}A(t,\mu)v=u_{n+1}t^{\alpha}A_{\lambda}(t,\mu)y(t), \tag{4.73a}$$

$$u_{n+1}=\text{const.}; \tag{4.73b}$$

$$B(\mu)v(1)+u_{n+1}(1)B_{\lambda}(\mu)y(1)=0, \tag{4.74a}$$

$$y(1)^T v(1)=0; \tag{4.74b}$$

$$v\in C([1,\infty]). \tag{4.75}$$

Because of Definition 4.1 the problem (4.73), (4.74), (4.75) has no solution unless $u_{n+1}=0$. If $u_{n+1}=0$ then v has to be an eigenfunction of (4.1), (4.2), (4.3); therefore $v=cy$ for some constant c. (4.74b) gives $c=0$, such that $u\equiv 0$ follows as the unique solution of (4.70), (4.71), (4.72). Therefore, we conclude from Markowich (1983) that the infinite problem (4.67), (4.68), (4.69) can be approximated by finite interval problems of the form

$$w_T'=t^{\alpha}f(t,w_T), \qquad 1\le t\le T, \tag{4.76}$$

$$b(w_T(1))=0, \tag{4.77}$$

$$S(w_T(T))=0 \tag{4.78}$$

where $w_T=(w_T^1,\cdots,w_T^n,w_T^{n+1})^T$; $(w_T^1,\cdots,w_T^n)$ is the approximation to the eigenvector y and w_T^{n+1} is the approximation to λ. The superscript T denotes transposition. The choice of $S:\mathbb{R}^{n+1}\to\mathbb{R}^{r_+}$ is explained in Markowich (1982b). The analysis given there shows that we can take

$$S(w_T(T))=S_p(w_T^{n+1}(T))\cdot\begin{pmatrix} w_T^1(T) \\ \vdots \\ w_T^n(T) \end{pmatrix}. \tag{4.79}$$

Markowich (1982b) has proved that the solution w_T is locally (around z) unique for T sufficiently large and that

$$\|z-w_T\|_{[1,T]}\le\text{const.}\|S(z(T))\|=\text{const.}\|S_p(\mu)y(T)\| \tag{4.80}$$

holds. So we get the order of convergence given in Theorem (4.1) with $s=1$, because the boundary condition (4.78) is equivalent to (3.21).

If (4.63) holds, the order of convergence is

$$|\mu_T-\mu|\leq \text{const.}\, T^{\gamma}\exp\Big(\Big(\max_{i=1(1)r_-}\operatorname{Re}\mu_i(\lambda)+\rho\Big)-a(T)\Big), \tag{4.81}$$

and the same is true for the normed eigenvectors.

5. Case studies. The first problem we treat is the so-called radial Schrödinger equation of the Kepler problem (see Jürgens and Rellich (1976, Chap. 3, par. 9) which is given by

$$-u''+\{(1+l)lr^{-2}-2cr^{-1}\}u=-\lambda u, \qquad 1\leq r<\infty \tag{5.1}$$

with $l\in\mathbb{N}\cup\{0\}$, where $\mathbb{N}$ is the set of positive integers, and $c\in\mathbb{R}$.

The transformation

$$y=(y_1,y_2)^T=(u,u')^T \tag{5.2}$$

takes (5.1) into the system

$$y'=\underbrace{\begin{bmatrix}0 & 1\\ (1+l)lr^{-2}-2cr^{-1}+\lambda & 0\end{bmatrix}}_{A(r,\lambda)}y, \qquad 1\leq r<\infty, \tag{5.3}$$

such that

$$A(\infty,\lambda)=\begin{bmatrix}0 & 1\\ \lambda & 0\end{bmatrix} \tag{5.4}$$

holds.

The parameter λ occurs linearly in (5.3), but we will construct the nonlinear asymptotic boundary condition $S_p(\lambda)$.

The Jordan form $J(\infty,\lambda)$ is

$$J(\infty,\lambda)=\begin{bmatrix}\sqrt{\lambda} & 0\\ 0 & -\sqrt{\lambda}\end{bmatrix}, \quad r_+=1, \quad r_-=1, \quad D_+=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \quad D_-=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}, \tag{5.5}$$

and therefore the set $\Omega=\{\lambda\in\mathbb{C}\,|\,\operatorname{Re}(\sqrt{\lambda})\neq 0\}$ is given by

$$\Omega=\mathbb{C}-\{\lambda|\operatorname{Re}\lambda\leq 0\}. \tag{5.6}$$

With appropriate boundary conditions of the form

$$a_1y_1(1)+a_2y_2(1)=0, \tag{5.7}$$

we conclude from §4 that every eigenvalue $\lambda\in\Omega$ of (5.3), (5.7) is isolated and that the dimension of the nullspace equals 1.

A complete analysis of the problem is given in Jürgens and Rellich (1976). They show that there is an infinite sequence of eigenvalues $\lambda^{(n)}$:

$$\lambda^{(n)}=c^2(l+1+n)^{-2} \quad \forall n\in N_0, \tag{5.8}$$

and the eigenfunctions $y^{(n)}$ are given by

$$y^{(n)}(r)\equiv\exp\Big(-\sqrt{\lambda^{(n)}}\,r\Big)r^{l+1}p_n(r)=\exp\Big(-\sqrt{\lambda^{(n)}}\,r\Big)r^{l+n+1}(1+O(r^{-1})) \tag{5.9}$$

because $p_n(r)$ is a polynomial in r of degree n. They assumed that $a_1=\sin\alpha$, $a_2=\cos\alpha$ with $\alpha\in[0,2\pi)$.

A straightforward calculation gives for $\lambda \in \Omega$

(5.10)
$$E(\lambda)=\begin{bmatrix} \sqrt{\lambda} & -\sqrt{\lambda} \\ \lambda & \lambda \end{bmatrix}, \quad E^{-1}(\lambda)=\begin{bmatrix} \frac{1}{2\sqrt{\lambda}} & \frac{1}{2\lambda} \\ \frac{1}{-2\sqrt{\lambda}} & \frac{1}{2\lambda} \end{bmatrix}, \quad P_+(\lambda)=\begin{bmatrix} \frac{1}{2} & \frac{1}{2\sqrt{\lambda}} \\ \frac{\sqrt{\lambda}}{2} & \frac{1}{2} \end{bmatrix}.$$

Therefore, we conclude (since $P_+(\lambda)=E(\lambda)D_+E^{-1}(\lambda)$)

(5.11) $$S_p(\lambda)=\left[\sqrt{\lambda}, 1\right].$$

The approximating problems have the form

(5.12) $$x_R'=A(r,\lambda)x_R, \qquad 1\le r\le R, \quad R\gg 1,$$

(5.13) $$[a_1,a_2]x_R(1)=0,$$

(5.14) $$\left[\sqrt{\lambda}, 1\right]x_R(R)=0 \Leftrightarrow \sqrt{\lambda}\, x_R^{(1)}(R)+x_R^{(2)}(R)=0, \quad x_R=\left(x_R^{(1)}, x_R^{(2)}\right)^T.$$

Since $A(r,\lambda)=\left[\begin{smallmatrix}0 & 1\\ \lambda & 0\end{smallmatrix}\right]+O(r^{-1})$ and since $w(R, \mathrm{Re}(\sqrt{\lambda}))\le \mathrm{const.}\, R^{-1}$, the convergence analysis given in §4 shows

(5.15) $$\max\left(\left|\lambda^{(n)}-\lambda_R^{(n)}\right|, \left\|y^{(n)}-x_R^{(n)}\right\|_{[1,R]}\right)\le ce^{-\sqrt{\lambda^{(n)}}R}R^{l+n-1}$$

because the $\lambda^{(n)}$'s are simple.

(5.15) follows by using $\psi_-(r,\lambda^{(n)})=[y^{(n)}(r), \frac{d}{dr}y^{(n)}(r)]^T$. Therefore the estimate (4.81), where ρ appears in the exponent, can be improved using (4.57). From Markowich (1982a) we conclude that every boundary condition $Sx_R(R)=0$, where S is independent of λ and where

(5.16) $$SE(\lambda)\begin{pmatrix}1\\0\end{pmatrix}\neq 0 \quad \text{in } \Omega$$

holds, leads to convergence of the order

(5.17)
$$\max\left(\left|\lambda^{(n)}-\lambda_R^{(n)}\right|, \left\|y^{(n)}-x_R^{(n)}\right\|_{[1,R]}\right) \le c_1\left\|y^{(n)}(R)\right\| w\left(R, \mathrm{Re}\left(\sqrt{\lambda}\right)\right)\le c\exp\left(-\sqrt{\lambda^{(n)}}R\right)R^{l+n}.$$

Setting $S=[s_1,s_2]$, $s_1,s_2\in\mathbb{C}$, (5.16) is fulfilled if and only if

(5.18) $$s_1+s_2\sqrt{\lambda}\neq 0$$

holds. For example, the natural boundary condition

(5.19) $$x_R^{(1)}(R)=0$$

satisfies (5.18) ($s_1=1$, $s_2=0$) and the order of convergence given by (5.17) differs only by one power of R from the order of convergence produced by the "optimal" boundary condition (5.14).

The second problem we consider is the Orr–Sommerfeld equation (see Ng and Reid (1980)) which governs the stability of a laminar boundary layer in a parallel flow approximation:

(5.20) $$\frac{1}{iR\alpha}\left(\frac{d^2}{dz^2}-\alpha^2\right)^2\phi-\left\{(U(z)-\lambda)\left(\frac{d}{dz^2}-\alpha^2\right)\phi-U''(z)\phi\right\}=0$$

where $\alpha>0$, $R>0$ is the Reynolds number, $U(z)$ is the velocity distribution satisfying

$$U(z)=1+F(z)e^{-\omega z^2}, \qquad \omega>0, \quad F\in C^2([0,\infty]) \tag{5.21}$$

such that $U(\infty)=1$, $U''(\infty)=0$ holds. $\phi(z)e^{i\alpha(x-\lambda t)}$ represents the disturbance stream function. The boundary conditions for the Orr–Sommerfeld problem are:

$$\phi(0)=\phi'(0)=\phi(\infty)=\phi'(\infty)=0. \tag{5.22}$$

The new variable

$$y=(\phi,\phi',\phi'',\phi''')^T \tag{5.23}$$

gives the linear eigenvalue problem

$$y'=\underbrace{\left\{\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\f_1(z)&0&f_2(z)&0\end{bmatrix}+\lambda\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\a&0&b&0\end{bmatrix}\right\}}_{A(z,\lambda)}y, \qquad 0\le z<\infty, \tag{5.24}$$

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\end{bmatrix}y(0)=0, \tag{5.25}$$

$$y\in C([0,\infty]) \tag{5.26}$$

where

$$f_1(z)=-\left(\alpha^4+i\alpha R\left(\alpha^2U(z)+U''(z)\right)\right), \tag{5.27a}$$

$$f_2(z)=2\alpha^2+i\alpha RU(z), \tag{5.27b}$$

$$a=i\alpha^3R, \tag{5.27c}$$

$$b=-i\alpha R \tag{5.27d}$$

hold. The eigenvalues of $A(\infty,\lambda)$ are

$$\begin{aligned}&\nu_1(\lambda)=\alpha, \qquad &&\nu_2(\lambda)=\left(\alpha^2+i\alpha R(1-\lambda)\right)^{1/2},\\ &\nu_3(\lambda)=-\alpha, \qquad &&\nu_4(\lambda)=-\left(\alpha^2+i\alpha R(1-\lambda)\right)^{1/2},\end{aligned} \tag{5.28}$$

so that $\mathrm{Re}\,\nu_1(\lambda)$, $\mathrm{Re}\,\nu_2(\lambda)>0$; $\mathrm{Re}\,\nu_3(\lambda)$, $\mathrm{Re}\,\nu_4(\lambda)<0$ for all $\lambda\in\Omega=\mathbb{C}-\{\lambda\,|\,\mathrm{Re}\,\lambda=1, \mathrm{Im}\,\lambda\le-\frac{\alpha}{R}\}$. All eigenvalues $\lambda\in\Omega$ of (5.24), (5.25), (5.26) are isolated and the null-spaces are at most two dimensional.

The approximating problems have the form

$$x_Z'=A(z,\lambda)x_Z, \qquad 0\le z\le Z, \tag{5.29}$$

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\end{bmatrix}x_Z(0)=0, \tag{5.30}$$

$$S(\lambda)x_Z(Z)=0 \tag{5.31}$$

where $x_Z=(x_Z^{(1)},x_Z^{(2)},x_Z^{(3)},x_Z^{(4)})^T$ holds and $\lambda\in\Omega$.

As for the first example, we calculate the optimal boundary condition $S(\lambda)=S_p(\lambda)$:

$$S_p(\lambda)=\begin{bmatrix}\alpha\nu_2(\lambda) & \alpha+\nu_2(\lambda) & 1 & 0\\ \alpha\nu_2(\lambda)(\nu_2(\lambda)+\alpha) & \nu_2^2(\lambda)+\alpha\nu_2(\lambda)+\alpha^2 & 0 & -1\end{bmatrix}. \tag{5.32}$$

Since $A(z,\lambda)=A(\lambda)+O(z^2e^{-\omega z^2})$ holds, we get from (4.81) for simple eigenvalues $\lambda=\mu\in\Omega$:

$$\begin{aligned}&\max\bigl(|\mu-\mu_Z|,\|y-x_Z\|_{[0,Z]}\bigr)\\ &\quad\le\text{const.}\,Z^2e^{-\omega Z^2}w\bigl(Z,\min(\alpha,\operatorname{Re}\nu_2(\mu))\bigr)\exp\bigl(\max(-\alpha,\operatorname{Re}\nu_4(\mu)+\rho)Z\bigr)\end{aligned}\tag{5.33}$$

where y, x_Z are the normed eigenfunctions. In the most interesting case $\alpha<|\operatorname{Re}\nu_4(\lambda)|<1$, the order of convergence is $Z^2\exp(-\omega Z^2-2(\alpha-\rho)Z)$, $\rho>0$ sufficiently small, while linear asymptotic boundary conditions like

$$x_z^1(Z)=x_z^2(Z)=0\tag{5.34}$$

achieve a slower order of convergence, namely $\exp(-2(\alpha-\rho)Z)$ (see Markowich (1982a)).

If the velocity profile fulfills

$$U(z)=1+F(z)e^{-\omega z},\qquad \omega>0,\quad F\in C^2([0,\infty]),\tag{5.35}$$

instead of (5.21), then the order of convergence reduces to $\exp(-(\omega+2(\alpha-\rho)z))$. Ng and Reid (1980) performed numerical experiments (using the optimal boundary condition (5.32)) with the Blasius velocity profile (fulfilling (5.21)) and with $U(z)=1-e^{-z}$. Their results confirm that the order of convergence is faster for the Blasius profile (see their Table I and II).

Grosch and Orszag (1977) performed computations using the linear boundary condition (5.34) and obtained numerically the indicated order of convergence. These computations show the superiority of the optimal (nonlinear) boundary condition (5.32) for the Orr–Sommerfeld problem.

REFERENCES

C. E. Grosch and S. A. Orszag (1977), *Numerical solution of problems in unbounded regions: Coordinate transforms*, J. Comp. Phys., 25, pp. 273–296.

F. 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. 46–61.

_____ (1980b), *An approximation method for boundary value problems on infinite intervals*, Computing, 24, pp. 227–239.

K. Jürgens and F. Rellich (1976), *Eigenwerttheorie gewöhnlicher Differentialgleichungen*, Springer-Verlag, Berlin, Heidelberg, New York.

H. B. Keller (1975), *Approximation methods for nonlinear problems with application to two point boundary value problems*, Math. Comp. 29, pp. 464–474.

_____ (1976), *Numerical Solution of Two Point Boundary Value Problems*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia.

M. Lentini and H. B. Keller (1980), *Boundary value problems on semi-infinite intervals and their numerical solution*, SIAM J. Numer. Anal., 17, pp. 577–804.

P. Markowich (1982a), *Eigenvalue problems in infinite intervals*, Math. Comp., to appear.

_____ (1982b), *A theory for the approximation of solutions of boundary value problems on infinite intervals*, this Journal, 13 (1982), pp. 484–513.

_____ (1983), *Analysis of boundary value problems on infinite intervals*, this Journal, 14, pp. 11–37.

B. S. Ng and W. H. Reid (1980), *On the numerical solution of the Orr–Sommerfeld problem. Asymptotic initial conditions for shooting method*. J. Comp. Phys., 38, pp. 275–293.

T. Kato (1966), *Perturbation Theory for Linear Operators*, Springer-Verlag, New York.

G. Vainikko (1976), *Funktionalanalysis der Diskretisierungsmethoden*, Teubner, Leipzig.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0004 $01.25/0

CONNECTION FOR WAVE MODULATION*

R. E. MEYER† AND J. F. PAINTER‡

Abstract. A new approach is described to the connection of wave amplitudes across the turning points and singular points of second-order, linear, analytic, ordinary differential equations which can describe the modulation of physical waves or oscillators. The general class of singular points thereby defined (§2) contains many irregular ones of greater complexity than have been accessible before; however, genuine coalescence of singular points is not here considered. The asymptotic connection formulae are shown to result directly from the branch structure of the singular point (§3), indeed, to a first approximation, they reflect merely the gross, local branch structure. The proof (§5) relates the local structure of the solutions at the singular point to the asymptotic wave structure by a limit process justified by bounds on the degree of irregularity of solution structure.

1. Introduction.

The semiclassical Schrödinger equation

$$\varepsilon^2 d^2 w/dz^2 + p\, w(z) = 0 \tag{1}$$

with small parameter ε and analytic coefficient function $p(z)$ is central to a vast class of oscillation and wave modulation problems in physics and other sciences. Particular interest, especially for scattering theory, attaches to the "WKB" problem of connecting the wave approximations to solutions across roots or singular points of $p(z)$. The following introduces a mathematical connection method which is simpler and more general than any advanced before [Zwaan 1929, Langer 1931, Painter and Meyer 1981]. Simplification and clarification of connection theory is, in fact, the whole objective of the study to be reported, and generalization was used only as a help towards it.

One reason why this objective has proved elusive over the generations may be that the general, second-order, linear differential equation, of which (1) is the normal form, encompasses too many disparate phenomena. The present study focuses on only those forms of (1) which are genuine Schrödinger equations in the sense that they can describe the modulation of physical waves or oscillators. This subclass is characterized in §2 in terms of its admissible (turning-point and) singular-point structure. To attempt only one step at a time, moreover, genuine coalescence of singular points is excluded. This leaves a large class of singular points, none the less, because the potential functions $p(z)$ of (1) in the sciences must be defined, if not by speculation, then by measurement, in which case they can be known only imperfectly. In addition, it has long been recognized that scattering matrices may depend decisively on singular points of $p(z)$ away from the real axis of time or space, in which case there are scant physical grounds for restricting their nastiness. The characterization of $p(z)$ cannot therefore be very specific and must include *arbitrarily irregular* points of (1) in the sense that "rank" or similar notions of degree of irregularity are inapplicable. Certainly, multivalued functions $p(z)$ must be the norm, rather than the exception. All the same, modulation implies a certain structure (§2).

The multivaluedness of $p(z)$ in (1) implies that the solutions $w(z)$ must normally be multivalued, and the main thesis to be propounded is that this multivaluedness is the source of the connection problem and that the asymptotic connection formulae solving

* Received by the editors December 4, 1981, and in revised form May 4, 1982. This work was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and supported partially by the National Science Foundation under grants MCS-7700097 and MCS-8001960 and by the Wisconsin Alumni Research Foundation.

‡ Lawrence Livermore Laboratory, Livermore, California 94550.
† Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

it are a direct manifestation of the branch structure of the potential $p(z)$ at the singular point. It is not a new thesis, the same fundamental view of connection was adopted already by Olver [1974, pp. 481, 482] for regular and isolated singular points. Our objective is to show how it can be extended to large classes of very irregular ones and what new insights into the nature of connection emerge therefrom.

To this end, local solution representations near the singular point have been developed in a companion paper [Meyer and Painter 1982], hereafter referred to as [IPM], and are summarized in §3. They focus on a particular fundamental system y_s, y_m of (1) in which $y_m/y_s \to 0$ as $x \to 0$ so that y_m has a milder singularity than y_s which, in turn, contains no additive multiple of y_m. This makes the pair a characteristic representation of the branch structure of the singular point. The representation is a local one, in the first place, but turns out to have a striking two-variable structure: in the framework of a natural, independent variable x (§2), the solutions do not depend on the complex parameter ε in (1) separately, but only on the variables x and εx. One key property of the representations in regard to connection is that they are partially global in the oscillation variable x, even though strictly local in the modulation variable εx. Another, is that they can be extended to bounds on 'degrees of irregularity' of solution structure. For instance, regularity of a singular point implies certain global symmetries for the solutions and accordingly, an estimate of solution asymmetry and of its dependence on distance from the singular point constitutes a bound on a degree of irregularity. Again, any irregular point of the general kind here considered can be linked diffeomorphically to a regular point and a pointwise comparison of corresponding solutions of the differential equations so associated constitutes a 'diffeomorphic bound' on a degree of irregularity.

Such symmetry bounds are shown in §5 to admit a class of limit processes in which both $|x| \to \infty$ and $|\varepsilon x| \to 0$ and furthermore, the asymmetry of y_s and y_m is severely restricted. This amounts to showing that asymptotic approximation of WKB-type characterized by dominance and recessivity becomes available before local structure has been lost. Existence of such limit processes translates immediately local information on the structure of y_s and y_m at the singular point into information on the multivaluedness of their asymptotic wave-representation. But, the latter information is what connection theory seeks.

The simple symmetry characteristic of regular points fails for a nongeneric subset of 'Frobenius exceptions' [Olver 1974, p. 150] which involve logarithmic branch points, and for the corresponding, nongeneric subset of irregular points, the alternative approach to connection via diffeomorphic bounds is more helpful. It is used in §6 to continue the wave amplitudes analytically in the main connection parameter. While this continuation is documented only in an asymptotic sense more abstract than is really desirable, it does explain why no trace of the Frobenius exceptions is visible in the first approximation to the connection formulae and also illuminates the relation of Frobenius exceptions to solution normalization and poles of the gamma function. (More concrete, even if not quite as wide, coverage of Frobenius exceptions is provided by the connection method of Painter and Meyer [1982].)

To explain these facets of connection and irregularity, it is helpful to limit the presentation in other ways, and attention will be thus restricted to the first approximation to connection of wave amplitudes. A more exhaustive description may be thought desirable for the sake of completeness, and most of all, error bounds are desirable. It may be noted that the representations used are obtained by the standard method of Volterra integral equations, which is precisely the method leading to effective error bounds [Olver 1974]. The sketch of the representation method given in §3 makes clear

that such bounds would emerge in terms of pointwise and integral bounds on a certain irregularity function $\phi(\varepsilon x)$ arising in the characterization of the singular point (§2). In the general case, however, that function is barely specified and can tend to zero with εx arbitrarily slowly, so that it appears doubtful whether the bounds would give much satisfaction to the numerical analyst. Most of all, however, we suspect that the present proof of connection may not remain the simplest one for long, in which case present attention to error bounds and higher approximations may be premature.

2. Modulation equations. Equation (1),

$$\varepsilon^2 d^2w/dz^2+p(z;\varepsilon)w(z)=0,$$

is one of a large family of normal forms of the general, linear, second-order differential equation, and constructive general statements are difficult in such an indefinite frame. By constrast,

$$dx/dz=i\,p^{1/2}/\varepsilon$$

defines the Liouville–Green of WKB or Langer variables x and εx based on the local wavelength or period, which have long been recognized as the natural ones for wave modulation. Physical specifications, e.g. for scattering, relate directly to them, and if z differs substantially from x, it can at best measure distance in legal units such as inches or cm. The natural formulation of physical problems of wave modulation is therefore in terms of x or εx, from the start, which will avoid the extraneous difficulty of describing the global transformation between εx and z, which has no physical significance and can be a very complicated, multivalued map.

The exclusion of coalescence, in order to confine attention to one singular point at a time, restricts not their total number, but only how fast they can approach each other as $\varepsilon\to 0$. When this is not fast enough to introduce genuine coalescence, a rescaling [Meyer and Guay 1974] permits the elimination of the main ε-dependence from $p(z;\varepsilon)$, and therefore not much generality is lost by ignoring the residual ε-dependence. For simplicity, $p(z)$ will accordingly be taken independent of ε in what follows.

The main property distinguishing the wave modulation equations among the larger class of equations (1) is that the *natural variable x must be definable*, for otherwise, not even the concepts of wavelength or period could exist for (1). An additional requirement arises as follows. If $p^{1/2}$ be nonintegrable at a singular point, then that point is seen to correspond to no $x\in C$ and hence, represents not a genuine singularity but a device for reinterpreting a radiation condition as a singular point in the z-plane. Such a device has been used at times in quantum mechanics, but is excluded here to concentrate on the class of genuine singularities of modulation. For that class, *the singular point of* (1) *must correspond to a definite point x*. Without loss of generality, both will be identified with the origin.

For an effective formulation of this notion of the most general wave modulation equation (short of coalescence), it should be expressed in terms of the natural variable x. Accordingly, the following is based on the premise that a *branch $r(x)$ of $p^{1/4}$ is definable as an analytic function on a punctured neighborhood of $x=0$* which is a Riemann surface including the interval $(-\pi,2\pi)$ of $\arg x$ *so that*

$$i\,dz/dx=\varepsilon r^{-2} \tag{2}$$

is integrable at $x=0$. (In conventional, turning-point terminology, such a Riemann surface element comprises three adjacent Stokes sectors. Like $z(x)$, of course, $r(x)$ also depends parametrically on ε.)

When the Schrödinger equation (1) is transformed to the natural variable by, say, simply setting

$$w(z)=y(x),$$

it takes the form

$$y''+2r^{-1}r'y'-y=0, \tag{3}$$

which shows that the wave development is controlled by the "modulation function"

$$\frac{r'(x)}{r(x)}=\frac{i\varepsilon}{2}\frac{d}{dz}p^{-1/2}$$

rather than by the potential function $p(z)$ directly. This illuminates why it has long been known that the singular points of (1) should really include the roots of $p(z)$ ("turning points"). It is also seen that the modulation function has a particular structure: since $p(z)=r^4$ is a function of z independent of ε, it follows from (2) that εx is also such a function, and in turn, that xr'/r depends on x and ε only through the product εx. A secondary hypothesis to be now adopted, because it simplifies the theory of connection, is that a limit of xr'/r as $\varepsilon x\to 0$ can be identified,

$$xr'/r\to\gamma\in C \quad \text{as } \varepsilon x\to 0 \tag{4}$$

uniformly in the Riemann surface sector Δ of εx in which xr'/r is defined locally near $\varepsilon x=0$. These two hypotheses also define the framework of the analysis of [IPM]. A statement equivalent to (4) is that the (fourth root of the) potential $r(x)$ can be written in the form

$$p^{1/4}=r(x)=(\varepsilon x)^{\gamma}\rho(\varepsilon x) \tag{5}$$

where $\rho(\xi)$ is a function analytic on the Riemann surface element Δ with the property

$$(\xi/\rho)\,d\rho/d\xi=:\phi(\xi)\to 0 \quad \text{as } \xi\to 0 \tag{6}$$

uniformly in Δ, because $\phi(\varepsilon x)=xr'/r-\gamma$.

To make the structure of the theory more readily apparent, it will help to abbreviate the notation by the convention that a function symbol such as $g(x)$ will, as before, be always understood to denote a function of both x and ε. By contrast, a Greek symbol such as $\psi(\xi)$ will always denote a function of $\xi=\varepsilon x$ only. If such a function has the property (6),

$$(\xi/\psi)\,d\psi/d\xi\to 0 \quad \text{as } \xi\to 0$$

uniformly in Δ, then it will be called *mild*; it implies that $\psi(\xi)$ varies near $\xi=0$ less than any nonzero real power of ξ:

$$\forall\nu>0, \quad |\xi^{\nu}\psi^{\pm 1}|\to 0 \quad \text{as } \xi\to 0.$$

In particular, the limit γ postulated in (4) is thus seen from (5) to represent the exponent of the "*nearest power*" of x in the (fourth root of the) potential, and the basic integrability premise defining physical Schrödinger equations implies

$$\operatorname{Re}\gamma\le\frac{1}{2}.$$

The general class of singular points of Schrödinger equations thus defined includes very irregular ones, in addition to all the turning points covered in the literature

[Painter and Meyer 1982]. For Langer's [1931] class of fractional turning points,

$$z^{2\gamma/(2\gamma-1)}[p(z)]^{1/2}$$

is analytic and nonzero at $z=0$,

$$\phi(\xi)=\sum_{n=1}^{\infty}\gamma_n\xi^{(1-2\gamma)n}$$

and the solutions of (1) and (3) are approximable uniformly in terms of Bessel functions [Langer 1931, Olver 1977]. For other singular points, however, no simple, uniform approximands in terms of classical functions appear likely. Local approximations have been constructed in [IPM] to provide support for the present study even under the vague assumptions just sketched, which admit functions $\phi(\xi)$ of arbitrary multivaluedness and approaching zero as $\xi\to 0$ more slowly than any specific function. Similarly, the coefficient functions $p(z)$ in (1) here admitted can be of great complexity, especially when several irregular singular points are present, and a useful global description appears unlikely in the general case. Locally, however, the class of potential functions $p(z)$ can be described by

$$z^{-1}\int_0^z[p(t)/p(z)]^{1/2}dt\to 1-2\gamma \quad \text{as } z\to 0.$$

As indicated in the Introduction, the conceptual key to connection lies in the two-variable structure of the Schrödinger equation emerging from (2) to (4). It should not have surprised us as much as it did, for it is already apparent in (1) that the independent variable plays two distinct physical roles. The first term in (1) represents the oscillatory mechanism, and its independent variable is clearly z/ε, with local wavelength as natural unit, prompting the transformation to

$$x=\frac{i}{\varepsilon}\int_0^z\sqrt{p}\,(t)\,dt. \tag{7}$$

By contrast, $p(z)$ represents the potential and its variation, which is not dependent on the presence of waves, and the role of z in it is therefore a different one. The formulation sketched in this section adds the insight that the dependence on the modulation variable $z\leftrightarrow\varepsilon x=\xi$ enters into the Schrödinger equation (3) only through a relatively minor term

$$\phi(\varepsilon x)=xr'/r-\gamma$$

in the modulation function r'/r. This function ϕ has been called "irregularity function" in [IPM] because $\phi=0$ characterizes the regular singular points.

The critical role of this two-variable structure will emerge in §5 which is devoted to a proof that connection across singular points is a mathematical process *local* in εx, even though *asymptotic* in x. This is, perhaps, the main new insight gained by extending the fundamental view of connection of Olver [1974] to irregular singular points. It explains why a merely local definition of $\phi(\varepsilon x)$—and thereby, of the potential and of the Schrödinger equation—on a Riemann surface element

$$\Delta=\{\varepsilon x:\ -\pi\le\arg(\varepsilon x)\le 2\pi,\ 0<|\varepsilon x|<E \text{ for some } E>0\}$$

turns out sufficient. (For notational convenience, E is adjusted so that $\phi(\varepsilon x)$ is analytic up to the rim of the element and hence, *bounded* on Δ.) In a *shortwave limit* $\varepsilon\to 0$, that entails little restriction on the corresponding domain D of x. On account of the two-variable structure, moreover, a shortwave limit must be a limit $\varepsilon x\to 0$. Hindsight, of

course, makes all of this appear foreshadowed in the structure of (1), where the first, oscillation term is defined globally in z/ε, even if the potential $p(z)$ be defined only locally.

3. Branch structure. It will help to summarize now the results of [IPM] used in the proof of connection in §5 and to indicate their motivation. If a limit $\varepsilon \to 0$ is taken at fixed x, then $\phi(\varepsilon x) \to 0$, by (5), and (3) approaches a form of Bessel's equation (which observation started Langer's [1931] work). Its singular point is regular with Frobenius exponents 0 and $1-2\gamma \geq 0$. If $1-2\gamma$ is not an integer, that implies solutions $f_s(x)$ and $x^{1-2\gamma} f_m(x)$ with *entire* functions f_s, f_m (and $f_s(0)=f_m(0)=1$), which turn out to depend only on x^2. Integer values of $\frac{1}{2}-\gamma$ correspond to the Frobenius exceptions for which only $f_m(x)$ is entire.

It is plausible that (3) may have an analogous fundamental system (y_s, y_m) when $\varepsilon \neq 0$, which displays the branch structure of the irregular singular point most clearly. If $y_m/y_s \to 0$ as $x \to 0$, then y_s has there the stronger singularity and it appears natural to call it the stronger solution, and y_m, the milder. Such a system has been constructed [IPM] to obtain a representation of the branch structure at the general irregular point of wave modulation and, in particular, to find out what replaces the entire functions f_s, f_m and explore how departure from entirety can be characterized. The underlying idea emerges most simply in the following construction of $y_s(x)$ for $\frac{1}{2} \geq \mathrm{Re}\,\gamma > -\frac{1}{2}$.

Since (3) can be written $(r^2 y')' = r^2 y$, a simple Volterra equation associated with it is

$$
\begin{aligned}
y'(x) &= [r(x)]^{-2} \int_0^x [r(v)]^2 y(v)\, dv, \\
y(x) &= 1 + \int_0^x y'(v)\, dv.
\end{aligned}
\tag{8}
$$

By (5), a plausible iterative approach to (8) is by a sequence $\{b_n(x)\}$ such that

$$
\begin{aligned}
&\frac{db_{n+1}}{dx} = \int_0^x \left[\frac{\rho(\varepsilon v)}{\rho(\varepsilon x)}\right]^2 \left(\frac{v}{x}\right)^{2\gamma} b_n(v)\, dv, \\
&b_0 \equiv 1, \qquad b_n(0) = 0 \quad \text{for } n \geq 1
\end{aligned}
\tag{9}
$$

(and as always, it is understood that $b_n(x)$ depends also on ε). Since $\rho(\xi)$ is a mild function, it can be estimated in (9) at the expense of a small power, and it emerges readily in this way, by estimation of b_n' from (9) and (6) and thence, b_n, recursively, that

$$
\begin{aligned}
&b_n(x) = \beta_n(\varepsilon x)(x/2)^{2n}, \\
&|\beta_n(\xi)| \leq k_n' = \Gamma(-s)/[n!\,\Gamma(n-s)], \\
&s = -\mathrm{Re}\,\gamma - \frac{1}{2} + \operatorname*{lub}_{u \in (0,1)} |\phi(u\xi)|.
\end{aligned}
\tag{10}
$$

Therefore, if $|\varepsilon x| < E(\gamma)$ chosen to assure $s<0$, then the rapid decrease of k_n' with n documents a majorant series assuring the convergence of Σb_n to a solution $y_s(x)$ of (3) analytic on the Riemann sector D. Since $y_s(0)=1$, $\Sigma b_n(x)$ generalizes Frobenius' entire function $f_s(x)$, but the 'coefficients' β_n are generally multivalued functions of εx.

Observe that (10) suggests $y_s = \Sigma b_n$ tends to an even function of x in some sense as $\varepsilon x \to 0$. This can be made more precise by applying the same approach to the estimation

of $|b_n'(x)+b_n'(x\exp-\pi i)|$ and thence, $|b_n(x)-b_n(x\exp-\pi i)|$ to show [IPM] that

$$|b_n(x)-b_n(xe^{-\pi i})|\leq\delta_s(|\varepsilon x|)nk_n'|x/2|^{2n}$$

where

$$\delta_s(|\xi|)\to 0 \quad \text{as } |\xi|\to 0. \tag{11}$$

These bounds still decrease fast enough with n to be summed to a bound on the degree of oddness of $y_s(x)$ in terms of the modified Bessel function $I_\nu(z)$ [Olver 1974, p. 60]: For $\frac{1}{2}\geq\mathrm{Re}\,\gamma>-\frac{1}{2}$ and x, $x\exp(-\pi i)$ in D and $|\varepsilon x|<E(\gamma)$,

$$|y_s(x)-y_s(xe^{-\pi i})|\leq\delta_s(|\varepsilon x|)\Gamma(-s)|x/2|^{2+s}I_{-s}(|x|). \tag{12}$$

Thus $y_s(x)$ approaches evenness as $|\varepsilon x|\to 0$ uniformly on compact subsets of the cut x-plane. For fixed εx, on the other hand, $y_s(x)$ may lose evenness exponentially fast as $|x|$ increases.

A good representation of the milder solution $y_m(x)$ of (3) depends on identification of the exact function generalizing the factor $x^{1-2\gamma}$ of $f_m(x)$. It turns out to be just the function $z(x)$ defined by (2), indeed [IPM],

$$z(x)=(\varepsilon x)^{1-2\gamma}\zeta(\varepsilon x) \tag{13}$$

with a mild function $\zeta(\xi)$ such that

$$(1-2\gamma)\rho^2\zeta\to -i\varepsilon \quad \text{as } \xi\to 0. \tag{14}$$

Then $y_m(x)/z(x)=\hat{y}(x)$ satisfies a differential equation related to (3) and can be constructed for all $\mathrm{Re}\,\gamma\leq\frac{1}{2}$ by an iteration paralleling that just sketched to obtain [IPM] a representation

$$y_m(x)=z(x)\left[1+\sum_1^\infty\alpha_n(\varepsilon x)(x/2)^{2n}\right] \tag{15}$$

with bounds

$$|\alpha_n(\xi)|\leq k_n=\Gamma(m)/[n!\Gamma(m+n)],$$

where $m=\frac{3}{2}-\mathrm{Re}\,\gamma-\delta_2(|\varepsilon x|)>0$ for $|\varepsilon x|<$ another $E(\gamma)$, since $\delta_2(|\xi|)\to 0$ as $|\xi|\to 0$. Of course, $\zeta(\xi)$ and $\alpha_n(\xi)$ are generally multivalued functions, but $\alpha_n(0)$ is defined and nonzero, so that $y_m/z=\hat{y}(x)$ also tends to an even function as $\varepsilon x\to 0$. An oddness bound is obtained by an estimate paralleling that indicated above.

THEOREM 4[1] [IPM]. *For x and $xe^{-\pi i}$ in D and $|\varepsilon x|<E(\gamma)$,*

$$|\hat{y}(x)-\hat{y}(xe^{-\pi i})|\leq\delta_m(|\varepsilon x|)\Gamma(m)|x/2|^{2-m}I_m(|x|)$$

and $\delta_m(|\xi|)\to 0$ as $|\xi|\to 0$.

The same comment therefore applies to $y_m/z=\hat{y}$ as follows (12).

For the stronger solution y_s and for $\mathrm{Re}\,\gamma\leq-\frac{1}{2}$, the simple Volterra equation (8) can, by (5), be used only at the price of a regularization of the first integral in (8). A stronger solution in the sense indicated is defined only up to an additive multiple of the milder solution, which is undesirable for a fundamental system displaying clearly the branch structure of the singular point. The regularization adopted in [IPM] avoids that

[1] Theorems are numbered consecutively from [IPM].

additive multiple and constructs a stronger solution of the form

$$y_s(x)=1+\sum_1^\infty \beta_n(\varepsilon x)(x/2)^{2n} \tag{16}$$

for all $\operatorname{Re}\gamma\le\frac{1}{2}$, but the construction is different in different strips

$$S_N=\left\{\gamma\in\mathbb{C}: 1-N>\operatorname{Re}\gamma-\frac{1}{2}>-N\right\}, \qquad N=0,1,2,\cdots,$$

and the estimates of $|\beta_n|$ grow more laborious and weaker with increasing N. Since $\beta_n(0)$ is found to exist for all n only when $\frac{1}{2}-\gamma$ is not an integer, the symmetry even of the limit $\varepsilon x\to 0$ fails for integer $\frac{1}{2}-\gamma$. Within the strips S_N, however, symmetry bounds were established:

THEOREM 5[1] [IPM]. *For noninteger* $\frac{1}{2}-\operatorname{Re}\gamma\ge 0$, x *and* $xe^{-\pi i}$ *in* D *and sufficiently restricted* $|\varepsilon x|$,

$$|y_s(x)-y_s(xe^{-\pi i})|\le C(\gamma)\delta_s(|\varepsilon x|)|x/2|^{2+s}I_{-s}(|x|),$$

with $s=-\operatorname{Re}\gamma-\frac{1}{2}+\operatorname{lub}_{u\in(0,1)}|\phi(u\xi)|>0$ *and* $\delta_s(|\xi|)\to 0$ *as* $|\xi|\to 0$.

A useful light can be shed on the relation between the respective stronger solutions $y_{sN}(x)$ constructed in the strips S_N by reference to the regular point $\phi\equiv 0$ which the function $\phi(\varepsilon x)$ associates with any irregular point, by (5) and (6). Since (3) is a form of Bessel's equation for $\phi\equiv 0$, this open an avenue for pointwise comparison of, e.g., $y_{s1}(x)$ with an explicit function of x and γ:

THEOREM 6[1] [IPM]. *If* $\varepsilon x\in\Delta$, x *is bounded from* 0 *and* $\gamma\in$*compact* $G\subset T_1=\{\gamma\in\mathbb{C}: \frac{1}{2}\delta_\Delta>\operatorname{Re}\gamma>\delta_\Delta-\frac{1}{2}\}$, *then*

$$\left|y_{s1}(x)-\Gamma\left(\gamma+\frac{1}{2}\right)(x/2)^{(1/2)-\gamma}I_{\bar\gamma-1/2}(x)\right|\le C_1(G)\delta_2(\varepsilon x)e^{|x|}/|x^\gamma|$$

where $\delta_\Delta=\operatorname{lub}_{\xi\in\Delta}|\phi(\xi)|$ *and* $\delta_2(\xi)=\operatorname{lub}_{u\in(0,1)}|\phi(u\xi)|$.

Similar bounds [IPM], moreover, link y_{sN} on compact subsets of the strip

$$T_N=\left\{\gamma\in\mathbb{C}: \frac{3}{2}-N-\delta_\Delta>\operatorname{Re}\gamma>\frac{1}{2}-N+\delta_\Delta\right\}$$

with the same Bessel function, and by (6), all these stronger solutions are seen to approach, as $\varepsilon\to 0$ for fixed x, the same solution of (3) for $\varepsilon=0$.

4. Wave amplitudes. If $y(x)$ satisfies (3), then $W(x)=r(x)y(x)$ satisfies

$$W''=(1+r''/r)W, \tag{17}$$

and by (4)

$$r''/r=x^{-2}[\gamma(\gamma-1)+\phi(2\gamma-1+\phi+\xi\phi'/\phi)] \tag{18}$$

so that $|r''/r|$ is integrable along paths in D bounded away from $x=0$. This confirms [Olver 1974, p. 222] existence of a fundamental "WKB" solution pair

$$W_+(x)=a(x)e^x, \qquad W_-(x)=b(x)e^{-x} \tag{19}$$

with functions $a(x)$, $b(x)$ analytic on D and *bounded for large* $|x|$ (provided, of course, $\varepsilon x=\xi\in\Delta$ so that ϕ and $\xi\phi'$ are bounded). This is the fundamental system of (1) most strikingly describing the asymptotic wave character (undamped on the lines where x is pure imaginary) of the solutions.

The "amplitude functions" a, b are determined only up to a constant factor, but apart from that, the decay of $|r''/r|$ at large $|x|$ suffices [Olver 1974, p. 223, 224] to assure limits for $a(x)$ and $b(x)$ as $|x|\to\infty$ with $\arg x$ any integer multiple of π. Those limits are the well-known wave amplitudes in the first-order WKB-approximations to the solutions of (1).

Since any solution of (17) must be a linear combination of W_+ and W_-,

$$r(x)y_m(x)=\tilde{a}_m(x)e^x+\tilde{b}_m(x)e^{-x}, \tag{20}$$

and the same holds with subscript s in place of m, and supposing they can be normalized satisfactorily, then $\tilde{a}_m,\cdots,\tilde{b}_s$ are similarly analytic and bounded for large $|x|$. Since the left-hand side of (20) has been seen in §3 to be multivalued, not all of $\tilde{a}_s,\cdots,\tilde{b}_m$ can be entire, and symmetry makes it plausible that all of them will usually turn out to be multivalued. This prompts the question

$$\tilde{a}_m(\infty)-\tilde{a}_m(\infty e^{2\pi i})=?$$

which is, in fact, a connection question for WKB coefficients [Olver 1974, p. 481].

In view of the many contexts in which connection is important, it is natural that many different forms of the connection prblem are found in the literature, but most of them can be related to each other with little work, and a treatise on connection for simple turning points is found in [Olver 1974, Chap. 13]. In any case, the problem turns on relating the respective limits which represent the WKB coefficients on different domains, and when it is recognized that those domains correspond, in the frame of the natural variable, to sheets of the Riemann surface of the solution, then the form of the connection question just arrived at is seen to be a natural one.

By contrast to the functions $a(x)$, $b(x)$ first mentioned, $\tilde{a}_m(x)$ and $\tilde{b}_m(x)$ are normalized implicitly by the normalization of $y_m(x)/z(x)=\hat{y}(x)$, and this turns out to introduce an ε-dependence into the normalization of $\tilde{a}_m$, $\tilde{b}_m$. For fixed $\varepsilon\neq 0$, moreover, $|x|$ is bounded by E/ε on the Riemann sector D on which the differential equation (1) has been defined, and the connection question can therefore be posed only in the limit $\varepsilon\to 0$. This aspect is discussed in the Appendix, where it is shown that the functions

$$\varepsilon^{\gamma-1}\tilde{a}_m/(\rho\zeta)=a_m(x;\varepsilon x) \quad\text{and}\quad \varepsilon^{\gamma-1}\tilde{b}_m/(\rho\zeta)=b_m(x;\varepsilon x),$$

rather than $\tilde{a}_m$ and $\tilde{b}_m$ themselves, are certain to have limits as $\varepsilon\to 0$ and $|x|\to\infty$, and those limits are therefore the wave amplitudes of the milder solution. For an assuredly meaningful connection question, we should therefore rewrite the identity (20) as

$$\varepsilon^{\gamma-1}r(x)[\rho(\xi)\zeta(\xi)]^{-1}y_m(x)=a_m(x)e^x+b_m(x)e^{-x} \tag{21}$$

(with explicit mention of the dependence of a_m, b_m on $\xi=\varepsilon x$ omitted to focus attention now on $\arg x$) and ask $a_m(\infty e^{2\pi i})-a_m(\infty)=?$

5. Connection. Since (21) is an identity in x on D, it holds equally at $x\exp(-\pi i)$, if that point is also in D. If $\exp(-\pi i)$ be abbreviated by j, then since $y_m=z\hat{y}$ and $rz=x^{1-\gamma}\rho\zeta$, by (5) and (13), the identity

$$\begin{aligned}[\hat{y}(x)-\hat{y}(jx)]x^{1-\gamma}e^{-|x|}=&\left[a_m(x)-j^{\gamma-1}b_m(jx)\right]e^{x-|x|}\\&+\left[b_m(x)-j^{\gamma-1}a_m(jx)\right]e^{-x-|x|}\end{aligned} \tag{22}$$

also holds on D. Now let $|\varepsilon x|\to 0$, but $|x|\to\infty$ in such a way that the left-hand side of (22) still tends to zero. That this does indeed define a nonempty set of "intermediate limits", in the terminology of singular-perturbation theory [Eckhaus 1979], is a corollary of Theorem 4 because [Olver 1974, p. 435]

$$I_m(|x|)\sim|2\pi x|^{-1/2}e^{|x|}\quad\text{as }|x|\to\infty$$

and, e.g., $|x|=|\log\delta_m(|\varepsilon x|)|$ will serve.

For the choices $\arg x=\pi$ and $\arg x=2\pi$, respectively, such a limit of (22) yields

$$b_m(\infty e^{\pi i})=j^{\gamma-1}a_m(\infty),\qquad a_m(\infty e^{2\pi i})=j^{\gamma-1}b_m(\infty e^{\pi i}),\tag{23}$$

whence

$$a_m(\infty e^{2\pi i})-a_m(\infty)=2i\sin(\gamma\pi)\,b_m(\infty e^{\pi i}).\tag{24}$$

The choice $\arg x=0$ adds

$$a_m(\infty)=j^{\gamma-1}b_m(\infty e^{-\pi i})\tag{25}$$

to (23), whence the answer to the connection question for b_m is

$$b_m(\infty e^{\pi i})-b_m(\infty e^{-\pi i})=2i\sin(\gamma\pi)a_m(\infty).\tag{26}$$

For noninteger $\frac{1}{2}-\operatorname{Re}\gamma>0$, a parallel argument for the stronger solution starts from the identity

$$r(x)y_s(x)=\tilde a_s(x)e^x+\tilde b_s(x)e^{-x}\tag{27}$$

analogous to (20) to deduce that the normalization $y_s(0)=1$, $y_s'(0)=0$ assures boundedness of

$$\varepsilon^{-\gamma}\tilde a_s(x)/\rho(\xi)=a_s(x)\quad\text{and}\quad\varepsilon^{-\gamma}\tilde b_s(x)/\rho(\xi)=b_s(x)\tag{28}$$

as $\xi\to 0$ and leads via the identity (22) with $\hat y, m$ and $1-\gamma$ replaced, respectively, by y_s, s and γ, by the help of Theorem 5 to the same connection formulae (24), (26) for a_s, b_s in the place of a_m, b_m, because $\sin[(1-\gamma)\pi]=\sin(\gamma\pi)$. (It is this independence of subscript which makes (24), (26) more convenient for present purposes than various other relations, such as $a_m(\infty\exp 2\pi i)=a_m(\infty)\exp(-2\gamma\pi i)$, also implied by (23) and (25), or their counterparts for a_s, b_s obtained by replacement of $1-\gamma$ by γ.)

With appropriate interpretation, morever, the same connection fomulae relate $\tilde a_m$, $\tilde b_m$ and $\tilde a_s$, $\tilde b_s$, respectively, because $(\rho^2\zeta)^{-1}$ tends to a definite limit as $\xi\to 0$, by (14), and by (6),

$$\frac{\rho(j\xi)}{\rho(\xi)}=\exp\int_\xi^{j\xi}\phi(\tau)\,\frac{d\tau}{\tau}=\exp\int_0^{-\pi}\phi\big(|\xi|e^{i\sigma}\big)i\,d\sigma\to 1.$$

Since any solution $y(x)$ of (3) is a linear combination of the fundamental pair (y_s, y_m), the functions $a(x)$ and $b(x)$ in the representation

$$r(x)y(x)=a(x)e^x+b(x)e^{-x}$$

in terms of the fundamental pair (W_+, W_-) of (17) are linear combinations of $\tilde a_s$, $\tilde a_m$ and $\tilde b_s$, $\tilde b_m$, respectively, and therefore satisfy (24) and (26) as well. In the limit $\varepsilon\to 0$ and with interpretation appropriate to the normalization of $y(x)$, the connection formulae (24) and (26) will therefore be a general corollary of (3) under the two hypotheses of §2, when they have been extended in the next section to all γ with $\operatorname{Re}\gamma<\frac{1}{2}$.

6. Analytic continuation. Another proof of connection may be based on the diffeomorphic bounds, but will be detailed here only insofar as it covers the gaps left by the symmetry bounds, or at least, elucidates how the Frobenius exceptions fit, none the less, into connection theory.

By (27) and (28), the simple identity for the stronger solution corresponding to (21) is

$$x^{\gamma}y_s(x)=a_s(x)e^{x}+b_s(x)e^{-x}.$$

The wave amplitude $a_s(\infty)$ of the solution y_{s1} defined in the strip T_1 (§3), understood again as the limit as $\varepsilon\to 0$ and $|x|\to\infty$ with $\arg x=0$, is therefore

$$a_s(\infty)=\lim_{\substack{x\to+\infty\\ \varepsilon\to 0}}\left[e^{-x}x^{\gamma}y_{s1}(x)\right].$$

For fixed x and ε and given irregularity function $\phi(\varepsilon x)$ on Δ, the Schrödinger equation (3) depends analytically on γ, by (5), and therefore, so does y_{s1} in the domain T_1. If the limit defining the amplitude $a_s(\infty)$ is taken so that also $\varepsilon x\to 0$, as in §5, then (6) and Theorem 6 show $e^{-x}x^{\gamma}y_{s1}(x)$ to approach a limit uniformly on compacts $\subset T_1$ and hence, also $a_s(\infty)$ depends there analytically on γ. In fact [Watson 1944, p. 203], this limit is $\pi^{-1/2}2^{\gamma-1}\Gamma(\gamma+\frac{1}{2})$ and furnishes the analytic continuation of $a_s(\infty)$ to the domain C_Γ of $\Gamma(\gamma+\frac{1}{2})$.

Analogous statements apply to the amplitudes $a_s(\infty\exp 2\pi i),\cdots b_s(\infty\exp -\pi i)$ of y_{s1} on T_1 and show all of them to have analytic continuations to C_Γ which, in turn, are there related by the analytic continuations of the connection formulae. As shown in §5, those agree on T_N with the connection formulae for y_{sN} (of which the wave amplitudes also have analytic continuations to C_Γ, by the estimates similar to Theorem 6 quoted in [IPM]). In this sense of analytic continuation to C_Γ, the amplitudes and connection formulae are therefore independent of N. The remaining poles of connection, finally, can be removed by renormalization of the stronger solution or by posing the connection question for amplitude ratios such as

$$\frac{a_s(\infty e^{2\pi i})-a_s(\infty)}{b_s(\infty e^{\pi i})}=2i\sin(\gamma\pi).$$

Appendix. The somewhat delicate issue of normalization for connection may be brought under control in two steps. For fixed ε, the domain D of x on which $r(x)$, and hence also the differential equation (17), is defined is a Riemann sector of radius $E(\gamma)/\varepsilon$. Let the particular functions a,b in (19) normalized in the manner of Olver [1974, pp. 194, 220–222] be denoted by a_0, b_0. Then $a_0[b_0]=1$ and $a_0'[b_0']=0$ at a point of $\min_D[\max_D]\mathrm{Re}\,x$, which depends on ε, and thus $a_0=a_0(x;\varepsilon)$, $b_0=b_0(x;\varepsilon)$, and the first step will be to confirm that their dependence on ε weakens as $\varepsilon\to 0$.

Since $\phi(\xi)$ is analytic and bounded on the Riemann sector Δ for $0<|\xi|<E(\gamma)$, the same follows for

$$x^2\psi(x)=\gamma(\gamma-1)+(2\gamma-1)\phi(\xi)+\phi^2+\xi\phi'(\xi)=x^2r''/r$$

in (18) because it tends to $\gamma(\gamma-1)$ as $\xi=\varepsilon x\to 0$, by (6). For the basic connection question of §4, it is sufficient to restrict the Riemann sector D to a disc of the same radius cut along the positive real axis of x for $a_0(x;\varepsilon)$, and along the negative real axis, for $b_0(x;\varepsilon)$. Olver's [1974, p. 221] variation function for a_0 is

$$\Psi(x)=\int_{-|E/\varepsilon|}^{x}|\psi(v)\,dv|$$

evaluated along progressive paths, and if such a path keeps distance R from the origin, then since $x^2\psi$ is bounded,

(A1) $$\Psi = O(R^{-1}) \quad \text{as } R \to \infty.$$

The same holds for the variation function for b_0, which differs only in that the lower limit is $+|E/\varepsilon|$. The functions furnish [Olver 1974, p. 221] bounds

(A2) $$|a_0(x;\varepsilon)-1|,\ |a_0'(x;\varepsilon)| \le e^{\Psi(x)}-1 = O(R^{-1})$$

and similarly, for b_0. If now $|\varepsilon_k| < |\varepsilon_i|$, then $D(\varepsilon_k) \supset D(\varepsilon_i)$ and on $D(\varepsilon_i)$ the solution W_+ normalized for $D(\varepsilon_i)$ is a linear combination of W_+, W_- normalized for $D(\varepsilon_k)$; thus

$$a_0(x;\varepsilon_i)e^x = c_{ik}a_0(x;\varepsilon_k)e^x + d_{ik}b_0(x;\varepsilon_k)e^{-x}$$

with constants c_{ik}, d_{ik} computed from the respective normalizations and bounds to yield

$$a_0(x;\varepsilon_i) = a_0(x;\varepsilon_k)\{1+O(\varepsilon_i)\} + b_0(x;\varepsilon_k)e^{-2(x+E/\varepsilon_i)}O(\varepsilon_i)$$

and on $D(\varepsilon_i)$, $\mathrm{Re}(x+E/\varepsilon_i) \ge 0$. As long as $|x|$ is well bounded away from 0, therefore, $a_0(x;\varepsilon_i) = a_0(x;\varepsilon_k) + O(\varepsilon_i)$ as $|\varepsilon_i| \to 0$, and similarly for $b_0(x;\varepsilon)$, and by the bounds (A1), (A2), a_0 and b_0 tend to limits as $|x| \to \infty$ in $D(0)$.

For the second step, note that the amplitude functions $\tilde{a}_m$ and $\tilde{b}_m$ in (20) are renormalizations of a_0 and b_0 so that

$$\tilde{a}_m(x;\varepsilon) = A\,a_0(x;\varepsilon), \qquad \tilde{b}_m(x;\varepsilon) = B\,b_0(x;\varepsilon)$$

with coefficients A, B possibly dependent on ε. From (20), (5) and (13), therefore,

(A3) $$x^{1-\gamma}y_m(x)/z(x) = x^{1-\gamma}\hat{y}(x) = \varepsilon^{\gamma-1}(\rho\zeta)^{-1}\left[Aa_0(x;\varepsilon)e^x + Bb_0(x;\varepsilon)e^{-x}\right].$$

The differential equation for $\hat{y}(x)$ is, by (2) and (3),

$$\hat{y}'' + 2(r'/r + z'/z)\hat{y}' = \hat{y},$$

and since the normalization to $\hat{y}(0)=1$, $\hat{y}'(0)=0$ recognized in (15) is independent of ε, $\hat{y}$ inherits from $(r'/r+z'/z)x$ the property that it depends on ε only through $\xi = \varepsilon x$. Like $(r'/r+z'/z)x$, moreover, $\hat{y}$ depends continuously on ξ in Δ, for fixed $x \ne 0$, and as $\xi \to 0$, by (4) and (13), $r'/r + z'/z \to (1-\gamma)/x$ and the differential equation for $\hat{y}$ becomes a form of Bessel's, with solution

(A4) $$\lim_{\xi\to 0}\hat{y}(x) = \Gamma\left(\frac{3}{2}-\gamma\right)(x/2)^{\gamma-1/2}\,I_{(1/2)-\gamma}(x).$$

This shows $x^{1-\gamma}\hat{y}(x)$ to tend to a well-defined function of x on $D(0)$ as $\xi \to 0$, and since $a_0(x;\varepsilon)$ and $b_0(x;\varepsilon)$ have been shown to tend to limit functions on $D(0)$ as $\varepsilon \to 0$, it follows from (A3) that the functions $\varepsilon^{\gamma-1}A/(\rho\zeta)$ and $\varepsilon^{\gamma-1}B/(\rho\zeta)$ must tend to limits as $\varepsilon \to 0$ and $\xi \to 0$; of course, these limits might depend on the direction of approach, which is determined by $\arg x$.

In sum,

$$\frac{\tilde{a}_m}{\rho\zeta} = \frac{A}{\rho\zeta}a_0(x;\varepsilon), \qquad \frac{\tilde{b}_m}{\rho\zeta} = \frac{B}{\rho\zeta}b_0(x;\varepsilon),$$

where $\varepsilon^{\gamma-1}A/(\rho\zeta)$, $\varepsilon^{\gamma-1}B/(\rho\zeta)$ have limits as $\varepsilon \to 0$ and $\xi \to 0$, while the limits as $\varepsilon \to 0$ of a_0, b_0 are defined on $D(0)$ and tend there to limits as $|x| \to \infty$.

REFERENCES

W. ECKHAUS (1979), *Asymptotic Analysis of Singular Pertubations*, North-Holland, Amsterdam.

R. E. LANGER (1931), *On the asymptotic solutions of ordinary differential equations, with an application to Bessel functions of large order*, Trans. Amer. Math. Soc., 33, pp. 23–64.

R. E. MEYER AND E. J. GUAY (1974), *Adiabatic variation, Part* III, *A deep mirror model*, Z. Angew. Math. Phys., 25, pp. 643–650.

[IPM] R. E. MEYER AND J. F. PAINTER (1982), *Irregular points of modulation*, Adv. Appl. Math., to appear.

F. W. J. OLVER (1974), *Asymptotics and Special Functions*, Academic Press, New York.

_____, *Second-order differential equations with fractional transition points*, Trans. Amer. Math. Soc., 226, pp. 227–241.

J. F. PAINTER AND R. E. MEYER (1982), *Turning-point connection at close quarters*, this Journal, 13, pp. 541–544.

G. N. WATSON (1944), *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, London.

A. ZWAAN (1929), *Intensitaeten im Ca-Funkenspectrum*, Arch. Neerlandaises Sci. Exactes Natur. Ser. 3A, 12, pp. 1–76.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0005 $01.25/0

NONLINEAR INTEGRAL RICCATI SYSTEMS AND COMPARISON THEOREMS FOR LINEAR DIFFERENTIAL EQUATIONS*

G.J. BUTLER† AND L.H. ERBE†

Abstract. A generalization of a technique of Nehari [Trans. Amer. Math. Soc., 210 (1975), pp. 387–406] is introduced to study focal pairs for the pair of equations $L_n y+p(x)y=0$ and $L_n y+q(x)y=0$. Generalized Hille–Wintner theorems are obtained for the case when p, q are not necessarily of constant sign.

AMS-MOS subject classification (1980). Primary 34C10

1. Introduction. The application of nonlinear techniques to study problems in oscillation theory for linear ordinary differential equations is well known, and one of the most important methods, at least as far as the second-order case is concerned, involves the Riccati equation. In [8], Nehari extended these ideas to the study of nth order linear equations of the form

$$y^{(n)}+q(x)y=0 \tag{1.1}$$

by introducing associated Riccati differential systems. This leads to a variety of oscillation criteria for the equation. In this paper we wish to further develop these ideas in order to obtain comparison criteria for the more general equations

$$L_n y+p(x)y=0 \tag{1.2}$$

and

$$L_n y+q(x)y=0 \tag{1.3}$$

where p,q are continuous on an interval $I=(a,b)$ (or $(a,b],[a,b)$) with $a<b\leq+\infty$, and L_n is an nth order disconjugate linear differential operator on I (that is, the only solution of $L_n y=0$ with n zeros on I, counting multiplicities, is $y\equiv 0$). It is well known [10], [3] that L_n can be written in factored form as

$$L_n y=\rho_n(\rho_{n-1},\cdots,(\rho_1(\rho_0 y)')',\cdots)' \tag{1.4}$$

where $\rho_i>0$ and $\rho_i\in C^{n-i}(I)$. If we set $L_0 y=\rho_0 y$, $L_i y=\rho_i(L_{i-1}y)'$, $i=1,\cdots,n$, then $L_0 y, L_1 y,\cdots,L_n y$ are called the *quasiderivatives* of y (cf. [4]). (If $L_n y=y^{(n)}$, then $\rho_i\equiv 1$, $i=1,\cdots,n$, and the quasiderivatives are just the ordinary derivatives of y.) In several recent papers [3], [4], [5], Elias has studied the oscillatory character of (1.2) by means of a detailed analysis of the distribution of the zeros of the quasiderivatives. In these papers and in the papers of Nehari [8], [9], it was assumed that the coefficients $p(x)$ (and $q(x)$) were of constant sign on I. It will be shown that this requirement can be relaxed if one considers an appropriate Riccati integral system. We shall be primarily interested in theorems involving comparisons between the integrals of the coefficient functions. Typical of such theorems are those of Hille–Wintner type. Thus, if $n=2$ and $L_2 y=(r(x)y')'$, $r>0$, $I=(a,\infty)$, and if $0\leq|\int_x^\infty p(t)\,dt|\leq\int_x^\infty q(t)\,dt$, then disconjugacy of (1.3) on (a,∞) implies disconjugacy of (1.2) on (a,∞) (cf. [7], [11], and [2]). Generalizations to higher order equations, but with sign restrictions on the coefficients p,q, were obtained in [5] and [6].

*Received by the editors July 17, 1981, and in revised form April 22, 1982. This research was supported by grants from the Natural Sciences and Engineering Research Council, Canada.

†Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada, T6G 2G1.

We recall that an equation of the form (1.2) is said to be $(k,n-k)$ *disconjugate* on the interval $I=(a,b)$ in case there exists no nontrivial solution satisfying the boundary conditions

$$\begin{aligned} &L_i y(x_1)=0, \quad i=0,1,\cdots,k-1,\\ &L_j y(x_2)=0, \quad j=0,1,\cdots,n-k-1 \end{aligned} \tag{1.5}$$

for any x_1, $x_2\in I$ with $x_1<x_2$. Similarly, (1.2) is said to be $(k,n-k)$ *disfocal* on I in case there exists no nontrivial solution satisfying the conditions

$$\begin{aligned} &L_i y(x_1)=0, \quad i=0,1,\cdots,k-1,\\ &L_j y(x_2)=0, \quad j=k,\cdots,n-1 \end{aligned} \tag{1.6}$$

for any x_1, $x_2\in I$ with $x_1<x_2$. It was shown in [4] that if $I=(a,\infty)$ and p is of one sign, then (1.2) is $(k,n-k)$ disconjugate on I iff (1.2) is $(k,n-k)$ disfocal on I. It is easily seen (cf. [4],[8]) that (1.2) is a priori $(k,n-k)$ disconjugate (and $(k,n-k)$ disfocal) if $(-1)^{n-k}p>0$ on I. If (1.2) is $(k,n-k)$ disconjugate (or disfocal) on I for all $k=1,\cdots,n-1$, then (1.2) is disconjugate (resp. disfocal) on I. If not, then there will exist x_1, $x_2\in I$, $x_1<x_2$, and a nontrivial solution of (1.2) satisfying (1.5) (resp. (1.6)). More general boundary conditions than (1.5) or (1.6) may also be considered, and we refer the reader to [4] where extremal points are defined and their relationship to the oscillatory character of (1.2) is studied (see also [1]). For our purposes here, we shall have occasion to consider only focal-type conditions, that is, conditions (1.6) and

$$\begin{aligned} &L_i y(x_1)=0, \quad i=k,\cdots,n-1,\\ &L_j y(x_2)=0, \quad j=0,\cdots,k-1, \end{aligned} \tag{1.7}$$

where $x_1<x_2$. Conditions (1.6) and (1.7) motivate the following definitions. (The integer k is assumed fixed, $1\le k\le n-1$.) For fixed $x_1\in I$, the smallest $x_2\in I$, $x_2>x_1$, such that there exists a nontrivial solution of (1.2) satisfying (1.6) will be denoted by $\theta(x_1)\equiv x_2$. If no such x_2 exists, we set $\theta(x_1)=+\infty$. Similarly, for $x_1\in I$, the smallest $x_2\in I$, $x_2>x_1$, such that there exists a nontrivial solution of (1.2) satisfying (1.7) will be denoted by $\phi(x_1)\equiv x_2$, with $\phi(x_1)=+\infty$ if no such x_2 exists. Likewise, for $x_2\in I$, the largest $x_1\in I$, $x_1<x_2$, such that (1.2) has a nontrivial solution satisfying (1.6) will be denoted by $\hat{\theta}(x_2)\equiv x_1$ with $\hat{\theta}(x_2)=-\infty$ if no such x_1 exists. Finally, for $x_2\in I$, the largest $x_1\in I$, $x_1<x_2$, such that (1.2) has a nontrivial solution satisfying (1.7) is denoted by $\hat{\phi}(x_2)\equiv x_1$. It follows from the above definitions that if $\theta(x_1)=x_2$ then $\hat{\theta}(x_2)\ge x_1$, and if $\hat{\theta}(x_2)=x_1$ then $\theta(x_1)\le x_2$, with similar relations holding between ϕ and $\hat{\phi}$.

The systems technique which we develop here is motivated by a technique of Nehari [8] in which he studied $(k,n-k)$ disfocality (with sign assumptions on $p(x)$). We obtain comparison criteria for focal pairs (stated in terms of the functions defined above) and, along the way, correct a slight error in an argument in [8, Thm. 5.1]. The statements of the main results are given in §2. The proofs along with some additional technical lemmas are given in §3.

2. Main results. In our first result below we shall be interested in obtaining a comparison between focal pairs for (1.2) and (1.3) on the interval $I=(a,b]$, $a<b<+\infty$, corresponding to conditions (1.6). We shall denote these by $\hat{\theta}(p;b)$ and $\hat{\theta}(q;b)$, respectively. For convenience, we let

$$\hat{q}(x)\equiv(-1)^{n-k-1}\frac{q(x)}{\rho_0(x)\rho_n(x)}, \qquad \hat{p}(x)\equiv(-1)^{n-k-1}\frac{p(x)}{\rho_0(x)\rho_n(x)}.$$

THEOREM 2.1. *Assume* $\hat{q}(b)>0$ *and that*

$$\left|\int_x^b \hat{p}(t)\,dt\right| \le \int_x^b \hat{q}(t)\,dt, \qquad a<x\le b. \tag{2.1}$$

Then

$$\hat{\theta}(p;b)\le\hat{\theta}(q;b).$$

Similarly, we have a result for focal pairs corresponding to conditions (1.7) on $I=[a,b)$, $a<b\le+\infty$.

THEOREM 2.2. *Assume* $\hat{q}(a)>0$ *and that*

$$\left|\int_a^x \hat{p}(t)\,dt\right| \le \int_a^x \hat{q}(t)\,dt, \qquad a\le x<b. \tag{2.2}$$

Then

$$\phi(p;a)\ge\phi(q;a).$$

Using a result on the monotonicity of $(k,n-k)$ focal points (cf. [4]) under an assumption on the sign of the coefficients, we may state the following:

COROLLARY 2.3. *Assume* $(-1)^{n-k}p(x)<0$ *on* I *and let the assumptions of Theorem* 2.1 (*resp. Theorem* 2.2) *hold. Then* (1.2) *is* $(k,n-k)$ *disfocal on the interval* $(\hat{\theta}(q;b),b]$ (*resp.* $[a,\phi(q;a))$).

Theorem 2.2 applies to the infinite interval $I=(a,\infty)$ (or $[a,\infty)$) as well as the finite interval. The next result is an analogue of Theorem 2.1 for the case $I=(a,\infty)$.

THEOREM 2.4. *Assume* $(-1)^{n-k}p(x)<0$ *on* $I=(a,\infty)$ *and that* (1.3) *is* $(k,n-k)$ *disfocal on* I. *Assume*

$$0<\int_x^\infty \hat{p}(t)\,dt\le\int_x^\infty \hat{q}(t)\,dt, \qquad x\in(a,\infty).$$

Then (1.2) *is* $(k,n-k)$ *disfocal on* I.

3. Proof of Theorem 2.1. We define $w_j(x)$ to be the solution of (1.3) satisfying

$$L_{i-1}w_j(b)=\delta_{ij}, \qquad i,j=1,\cdots,n, \tag{3.1}$$

where $\delta_{ij}=1$ if $i=j$ and $=0$ if $i\neq j$. We define the k-dimensional row vector

$$U_i=(L_{i-1}w_1,L_{i-1}w_2,\cdots,L_{i-1}w_k), \qquad i=1,\cdots,n, \tag{3.2}$$

and let B denote the $k\times k$ matrix

$$B=\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_k \end{pmatrix}. \tag{3.3}$$

For convenience, set $\hat{a}\equiv\hat{\theta}(q;b)$. Then we claim that B is nonsingular on $(\hat{a},b]$. Notice first that $B(b)=I_k$, the $k\times k$ identity matrix and if $\det B(x_0)=0$ for some $\hat{a}<x_0<b$, then a nontrivial linear combination of $w_1,\cdots,w_k$ will have a kth order zero at x_0 and the quasiderivatives of order $k,k+1,\cdots,n-1$ will, by virtue of (3.1), vanish at b, contradicting the definition of $\hat{a}$.

Next we define the row vectors S_i by

$$S_i=(-1)^{k+i-1}U_iB^{-1}, \qquad i=k+1,\cdots,n. \tag{3.4}$$

We have

$$B' = \begin{pmatrix} \rho_1^{-1}U_2 \\ \rho_2^{-1}U_3 \\ \vdots \\ \rho_k^{-1}U_{k+1} \end{pmatrix} = CB + \rho_k^{-1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ U_{k+1} \end{pmatrix}$$

where C is the $k \times k$ matrix

$$C = \begin{pmatrix} 0 & \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & 0 & \rho_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdot & \cdots & \rho_{k-1}^{-1} \\ 0 & 0 & \cdot & \cdots & 0 \end{pmatrix}.$$

Differentiating the S_i's, we obtain

$$S_i' = -\frac{1}{\rho_i} S_{i+1} - (-1)^{k+i-1} U_i B^{-1} \left(C + \rho_k^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ U_{k+1} \end{pmatrix} B^{-1} \right)$$

$$= -\frac{1}{\rho_i} S_{i+1} - S_i \left(C + \rho_k^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ U_{k+1} \end{pmatrix} B^{-1} \right), \qquad i = k+1, \cdots, n-1,$$

and

$$S_n' = (-1)^{k+n-1} U_n' B^{-1} - (-1)^{k+n-1} U_n B^{-1} B' B^{-1}$$

$$= -\hat{q} U_1 B^{-1} - S_n \left(C + \rho_k^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ U_{k+1} \end{pmatrix} B^{-1} \right)$$

(since $U_n' = -(q/\rho_0\rho_n)U_1$). Since $U_1 B^{-1} = (1, 0, \cdots, 0)$ and since

$$S_i \begin{pmatrix} 0 \\ \vdots \\ 0 \\ U_{k+1} \end{pmatrix} = S_i^{(k)} U_{k+1},$$

where $S_i^{(k)}$ denotes the last (kth) component of the row vector S_i, we see that the S_i's satisfy the following system on $(a,b]$:

$$\begin{aligned} S_i' &= -\rho_i^{-1} S_{i+1} - S_i C - \rho_k^{-1} S_i^{(k)} S_{k+1}, \qquad i = k+1, \cdots, n-1, \\ S_n' &= -S_n C - \rho_k^{-1} S_n^{(k)} S_{k+1} - \hat{q}(1, 0, \cdots, 0), \end{aligned} \tag{3.5}$$

with $S_i(b) = 0$, $i = k+1, \cdots, n$.

If we let $\hat{e}$ denote the unit vector $\hat{e}=(1,0,\cdots,0)$, then (3.5) is equivalent to the integral system

$$
\begin{aligned}
S_i(x)&=\int_x^b\left(\rho_i^{-1}S_{i+1}+S_iC+\rho_k^{-1}S_i^{(k)}S_{k+1}\right)dt, \qquad i=k+1,\cdots,n-1,\\
S_n(x)&=\int_x^b\left(S_nC+\rho_k^{-1}S_n^{(k)}S_{k+1}\right)dt+\left(\int_x^b\hat{q}\,dt\right)\hat{e}.
\end{aligned}
\tag{3.6}
$$

Therefore, the hypotheses of the theorem imply that (3.6) has a continuous solution $S_{k+1},\cdots,S_n$ on $(\hat{a},b]$. Conversely, if there exists a continuous solution $S_{k+1},\cdots,S_n$ on $(a_1,b]$ for some $a\le a_1<b$, then we claim that B, as defined by (3.3), is nonsingular on $(a_1,b]$ and consequently $\hat{\theta}(q;b)\le a_1$. Suppose to the contrary that B becomes singular as $x\to x_0+$, where $a_1<x_0<b$. Then since $S_iB=(-1)^{k+i-1}U_i$ for $x_0<x\le b$, $i=k+1,\cdots,n$, it follows that $U_1,\cdots,U_n$ are contained in the space V spanned by $U_1,\cdots,U_k$. Moreover, since $\det B(x_0)=0$, it follows that $\dim V<k$ for $x=x_0$ and therefore there exists a nonzero constant vector $\alpha=(\alpha_1,\cdots,\alpha_k)$ which is orthogonal to V at $x=x_0$. That is, $\alpha\cdot U_i(x_0)=0$ (dot product) for all $i=1,\cdots,n$. If we set $w=\sum_{i=1}^n\alpha_iw_i$ (where the w_i's are as defined in (3.1)), then since $\alpha\cdot U_i(x_0)=L_{i-1}w(x_0)=0$, $i=1,\cdots,n$, we conclude that $w(x)\equiv0$, contradicting $\alpha\neq(0,\cdots,0)$. Therefore, B is nonsingular on $(a_1,b]$, and it is clear that $\hat{\theta}(q;b)\le a_1$.

We next observe (by Lemma 3.1 below) that the S_i's are all nonnegative on $(\hat{a},b]$. That is, the components $S_i^{(l)}\ge0$ on $(\hat{a},b]$ for $l=1,\cdots,k$ and $i=k+1,\cdots,n$. Next we consider the system

$$
\begin{aligned}
\sigma_i(x)&=\int_x^b\left(\rho_i^{-1}\sigma_{i+1}+\sigma_iC+\rho_k^{-1}\sigma_i^{(k)}\sigma_{k+1}\right)dt, \qquad i=k+1,\cdots,n-1,\\
\sigma_n(x)&=\int_x^b\left(\sigma_nC+\rho_k^{-1}\sigma_n^{(k)}\sigma_{k+1}\right)dt+\left(\int_x^b\hat{p}(t)\,dt\right)\hat{e},
\end{aligned}
\tag{3.7}
$$

which we may write as

$$
\tilde{\sigma}=T\tilde{\sigma} \tag{3.8}
$$

where $\tilde{\sigma}$ is the $(n-k)\times k$ matrix

$$
\tilde{\sigma}=\begin{pmatrix}\sigma_{k+1}\\ \sigma_{k+2}\\ \vdots\\ \sigma_n\end{pmatrix}. \tag{3.9}
$$

Consistent with our notation for the components $S_i^{(j)}$, we shall use the notation $\tilde{\sigma}_i^{(j)}$ for the (i,j)th entry of $\tilde{\sigma}$.

Let $\mathcal{C}$ be the Fréchet space of continuous $(n-k)\times k$ matrix-valued functions on $(a,b]$ with the compact-open topology and let $\mathcal{D}$ be the subset of $\mathcal{C}$ of functions $\tilde{\sigma}$ for which $|\sigma_i^{(j)}(x)|\le S_i^{(j)}(x)$, $a<x\le b$, $i=k+1,\cdots,n, j=1,\cdots,k$.

$\mathcal{D}$ is a closed convex subset of $\mathcal{C}$.

If $\tilde{\sigma}\in\mathcal{D}$, it follows from (3.7), that the function $(T\tilde{\sigma})(x)$ is continuous on $(a,b]$ and for $k+1\le i\le n-1$, $1\le j\le k$, we have

$$
\left|(T\tilde{\sigma})_i^{(j)}(x)\right|\le\int_x^b\left[\rho_i^{-1}S_{i+1}^{(j)}+(S_iC)^{(j)}+\rho_k^{-1}S_i^{(k)}S_{k+1}^{(j)}\right]dt=S_i^{(j)}(x).
$$

Similarly, $|(T\tilde{\sigma})_n^{(j)}(x)|\le S_n^{(j)}(x)$, $1\le j\le k$. Thus T maps $\mathcal{D}$ into itself, and the functions in the range $T(\mathcal{D})$ of T are uniformly bounded on compact subintervals of $(a,b]$.

The majorization of the components of members of $\mathfrak{D}$ by corresponding components $S_i^{(J)}$ and the integral form of the map T yields, in a standard fashion, the result that $T(\mathfrak{D})$ is equicontinuous on compact subintervals of $(a,b]$. Therefore $T(\mathfrak{D})$ is compact.

Since the components of $T\tilde{\sigma}$ are integrals of polynomial functions of the components of $\tilde{\sigma}$, it follows that T is a continuous mapping of $\mathfrak{D}$ into itself.

Now we may apply Tikhonov's theorem to deduce that T has a fixed point in $\mathfrak{D}$ which is a solution $\hat{\sigma}(x)$ of (3.7) on $(\hat{a},b]$. It follows therefore (as in the first part of the proof) that $\hat{\theta}(p;b)\le a$, and this completes the proof. □

LEMMA 3.1. *With the notation of Theorem* 2.1, *if* $S_i(x)=(S_i^{(1)}(x),\cdots,S_i^{(k)}(x))$, $i=k+1,\cdots,n$, *are solutions of the system*

$$
\begin{aligned}
S_i(x)&=\int_x^b\left(\rho_i^{-1}S_{i+1}+S_iC+\rho_k^{-1}S_i^{(k)}S_{k+1}\right)dt, \qquad i=k+1,\cdots,n-1,\\
S_n(x)&=\int_x^b\left(S_nC+\rho_k^{-1}S_n^{(k)}S_{k+1}\right)dt+\left(\int_x^b\hat{q}\,dt\right)\hat{e}
\end{aligned}
\tag{3.6}
$$

on $(\hat{a},b]$, *with* $S_i(b)=0$, $i=k+1,\cdots,n$, *then* $S_i^{(l)}(x)>0$ *on* $(\hat{a},b)$ *for* $i=k+1,\cdots,n$, $l=1,\cdots,k$.

Proof. Writing (3.6) in component form, we have

(3.10)

$$
\begin{aligned}
&S_i^{(1)}(x)=\int_x^b\left(\rho_i^{-1}S_{i+1}^{(1)}+\rho_k^{-1}S_i^{(k)}S_{k+1}^{(1)}\right)dt, \qquad i=k+1,\cdots,n-1,\\
&S_i^{(l)}(x)=\int_x^b\left(\rho_i^{-1}S_{i+1}^{(l)}+\rho_{l-1}^{-1}S_i^{(l-1)}+\rho_k^{-1}S_i^{(k)}S_{k+1}^{(l)}\right)dt, \quad i=k+1,\cdots,n-1,\ \ l=2,\cdots,k,\\
&S_n^{(1)}(x)=\int_x^b\rho_k^{-1}S_n^{(k)}S_{k+1}^{(1)}\,dt+\int_x^b\hat{q}\,dt,\\
&S_n^{(l)}(x)=\int_x^b\left(\rho_{l-1}^{-1}S_n^{(l-1)}+\rho_k^{-1}S_n^{(k)}S_{k+1}^{(l)}\right)dt, \qquad l=2,\cdots,k.
\end{aligned}
$$

We adopt the notation $f(x)\approx g(x)$ for functions f,g, if there exist positive constants A,B and an interval $J=(b-\delta,b)$ $(\delta>0)$ such that

$$Ag(x)\le f(x)\le Bg(x) \quad \text{for all } x\in J.$$

The symbols O and o will have their usual connotation.

We shall show that for $i=k+1,\cdots,n$, $l=1,\cdots,k$,

$$S_i^{(l)}(x)\approx(b-x)^{n+l-i}. \tag{3.11}$$

First we note that, since each component $S_i^{(l)}(x)$ is continuously differentiable and satisfies $S_i^{(l)}(b)=0$, we have

$$S_i^{(l)}(x)=O(b-x). \tag{3.12}$$

Let

$$\bar{S}_i^{(l)}(x)=\max_{x\le t\le b}S_i^{(l)}(t).$$

From the fourth equation of (3.10) and (3.12), we have

$$\bar{S}_n^{(l)}(x)=O\left((b-x)\bar{S}_n^{(l-1)}(x)\right)+O\left((b-x)^2\bar{S}_n^{(k)}(x)\right),$$

from which we obtain

$$\bar{S}_n^{(k)}(x)=O\big((b-x)\bar{S}_n^{(k-1)}(x)\big).$$

Successive use of (3.10) leads to

$$\bar{S}_n^{(k)}(x)=O\big((b-x)^{k-1}\big),$$

and so

$$S_n^{(k)}(x)=O\big((b-x)^{k-1}\big). \tag{3.13}$$

From (3.12), (3.13) and the third equation of (3.10), and noting the hypothesis $\hat{q}(b)>0$, we find that

$$S_n^{(1)}(x)\approx b-x,$$

which gives us (3.11) for $i=n$, $l=1$. Equation (3.13) and induction on l in the fourth equation of (3.10) give (3.11) for $i=n$, $l=1,\cdots,k$.

Assuming inductively that (3.11) holds for $i=m+1,\cdots,n$, $l=1,\cdots,k$, where $k+2\leq m+1\leq n$, we may use (3.13) and the second equation of (3.10) to eventually obtain

$$\bar{S}_m^{(k)}(x)=O\big((b-x)^{n+k-m}\big)+O\big((b-x)^{k-1}\bar{S}_m^{(1)}(x)\big). \tag{3.14}$$

Equation (3.14), the first equation of (3.10) and the inductive hypothesis give

$$\bar{S}_m^{(1)}(x)=O\big((b-x)^{n+1-m}\big), \tag{3.15}$$

and (3.14) and (3.15) give

$$\bar{S}_m^{(k)}(x)=O\big((b-x)^{n+k-m}\big). \tag{3.16}$$

Now the first equation of (3.10), (3.16) and the inductive hypothesis give

$$S_m^{(1)}(x)\approx(b-x)^{n+1-m},$$

which is (3.11) for $i=m$, $l=1$.

Then (3.16) and induction on l in the second equation of (3.10) give (3.11) for $i=m$, $l=1,2,\cdots,k$. Thus we have verified (3.11). It follows immediately that each component $S_i^{(l)}(x)>0$ on some interval $(b-\delta,b)$ where $\delta>0$.

Now it is easy to see from (3.6) that the components $S_i^{(l)}(x)$ stay positive throughout $(\hat{a},b)$, and the proof of the lemma is complete. □

Remark. In his proof of [8, Thm. 5.1] Nehari attempts to obtain positivity of the solution of a system somewhat analogous to (3.6). However, his argument for local positivity near the endpoint of the interval seems to be erroneous. An argument along the lines of the proof of the above lemma could be given instead.

Proof of Theorem 2.2. The proof of this theorem is very similar to the proof of Theorem 2.1 and can, in fact, be obtained by the change of variable $x-a=b-t$ along with appropriate modification in L_n. We leave the details to the interested reader. □

Proof of Corollary 2.3. Assume the hypotheses of Theorem 2.1 hold and that $(-1)^{n-k}p(x)<0$ on I. From results of Elias [4, Thm. 2] it follows that (in our terminology) $\theta(x)$ is a continuous nondecreasing function on its domain. Since $\hat{\theta}(p;b)\leq\hat{\theta}(q;b)\equiv\hat{a}$ we see that $\theta(p;\hat{a})\geq b$, and hence there cannot exist x_1, $x_2\in(\hat{a},b]$, $x_1<x_2$, and a nontrivial solution of (1.3) satisfying (1.6). Similarly, one obtains the conclusion of the corollary if the hypotheses of Theorem 2.2 hold. □

In order to prove Theorem 2.4, we state and prove the following technical lemma.

LEMMA 3.2. *Let $\{r_m\}$ be a sequence of real numbers with $r_m > a$ and $r_m \to \infty$, and suppose for each m that $g_m(x)$ is a nonnegative real-valued function defined on $a \le x \le r_m$ such that $g_m(x_1) \ge g_m(x_2)$ for $a < x_1 \le x_2 \le r_m$. Let $y_m(x)$ be a solution of the integral equation*

$$y_m(x) = g_m(x) + \int_x^{r_m} y_m^2(t)\,dt, \qquad y_m(r_m) = 0,$$

with $y_m(x) \equiv 0$ on $[r_m, \infty)$. Then $\{y_m(x)\}$ is uniformly bounded on each compact subinterval of (a, ∞).

Proof. Without loss of generality we may assume $a = 0$. Let $b > 0$ be given and let $\delta_1 = b/4$, $\delta_2 = b/2$. Then, assuming $r_m \ge b$, $y_m(x) \ge \int_x^{r_m} y_m^2(t)\,dt$ so that with $z_m(x) \equiv y_m(r_m - x)$ we have $z_m(x) \ge \int_0^x z_m^2(t)\,dt$. Define $v_m(x) \equiv \int_0^x z_m^2(t)\,dt$. Then, differentiating, we have $v_m' = z_m^2 \ge v_m^2$ and thus $v_m' v_m^{-2} \ge 1$. Integrating from $r_m - \delta_2$ to $r_m - \delta_1$ we obtain

$$\frac{1}{v_m(r_m - \delta_2)} - \frac{1}{v_m(r_m - \delta_1)} \ge \delta_2 - \delta_1,$$

and hence

$$\frac{1}{v_m(r_m - \delta_2)} \ge \delta_2 - \delta_1 + \frac{1}{v_m(r_m - \delta_1)} \ge \delta_2 - \delta_1.$$

Thus $v_m(r_m - \delta_2) \le 1/(\delta_2 - \delta_1)$ and so

$$\int_{\delta_2}^{r_m} y_m^2(t)\,dt \le \frac{1}{\delta_2 - \delta_1}.$$

It follows therefore that

$$\delta_2 \Big(\inf_{\delta_2 \le t \le 2\delta_2} y_m^2(t) \Big) \le \frac{1}{\delta_2 - \delta_1},$$

and so

$$\inf_{\delta_2 \le t \le 2\delta_2} y_m(t) \le \sqrt{\frac{1}{\delta_2(\delta_2 - \delta_1)}} \le \frac{4}{b}.$$

Thus, for all $x \in [2\delta_2, r_m] = [b, r_m]$ and $t \in [\delta_2, 2\delta_2]$, we have

$$y_m(x) = y_m(t) + g_m(x) - g_m(t) + \int_x^t y_m^2 \le y_m(t),$$

and hence

$$0 \le y_m(x) \le \inf_{\delta_2 \le t \le 2\delta_2} y_m(t) \le \frac{4}{b}.$$

If $r_m < b$, then $y_m(x) \equiv 0$ on $[b, \infty)$. It follows that $\{y_m(x)\}$ is uniformly bounded on $[b, c]$ for any $0 < b < c < \infty$. □

Remark. It is easy to see (and we shall make use of this fact) that if the equation in the hypotheses of the lemma is replaced by

$$y_m(x) = g_m(x) + \int_x^{r_m} h(t) y_m^2(t)\,dt$$

where $h(t)$ is a positive continuous function on (a, ∞), then the same conclusion holds.

Proof of Theorem 2.4. We choose a sequence of numbers $r_m \to \infty$ such that $a<r_m<\infty$ and such that $\int_x^{r_m}\hat{q}\,dt\geq 0$ and $\hat{q}(r_m)>0$. This we claim we can do because $\int_x^{\infty}\hat{q}\,dt>0$ for $x\in(a,\infty)$. That is, if it would not be possible to find such a sequence $\{r_m\}$, then for any sequence $\{r_m\}$ with $r_m\to\infty$ there would exist $x_m\in(a,r_m)$ such that we have $\int_x^{r_m}\hat{q}\,dt<0$. If we let, for each m,

$$\hat{x}_m=\inf\left\{x_m : a<x_m<r_m, \int_{x_m}^{r_m}\hat{q}\,dt<0\right\},$$

then we claim that $\{\hat{x}_m\}$ cannot be a bounded sequence. For if it were bounded, then choose a subsequence $\{\hat{x}_{m_k}\}$ which converges to $x_0\in(a,\infty)$. Then we would have (since $\int_{\hat{x}_m}^{r_m}\hat{q}\,dt\leq 0$)

$$0<\int_{x_0}^{\infty}\hat{q}\,dt=\lim_{k\to\infty}\int_{\hat{x}_{m_k}}^{r_{m_k}}\hat{q}\,dt\leq 0,$$

a contradiction. Thus, a subsequence of $\{\hat{x}_m\}$ which we again label $\{\hat{x}_m\}$ must diverge to ∞. But then for $a<x\leq\hat{x}_m$ we have

$$\int_x^{\hat{x}_m}\hat{q}\,dt=\int_x^{r_m}\hat{q}\,dt-\int_{\hat{x}_m}^{r_m}\hat{q}\,dt\geq 0$$

by definition of $\hat{x}_m$. This verifies the claim. Thus, for each m, from Theorem 2.1 we obtain a solution S_{im}, $i=k+1,\cdots,n$, of

$$(3.6)_m\qquad \begin{aligned} S_{im}(x)&=\int_x^{r_m}\left(\rho_i^{-1}S_{i+1,m}+S_{im}C+\rho_k^{-1}S_{im}^{(k)}S_{k+1,m}\right)dt, \qquad i=k+1,\cdots,n-1,\\ S_{nm}(x)&=\int_x^{r_m}\left(S_{nm}C+\rho_k^{-1}S_{nm}^{(k)}S_{k+1,m}\right)dt+\left(\int_x^{r_m}\hat{q}\,dt\right)\hat{e} \end{aligned}$$

which exists on $(a,r_m]$.

We next wish to proceed to a solution of the integral system on (a,∞) which corresponds to (3.6), i.e.,

$$(3.17)\qquad \begin{aligned} S_i(x)&=\int_x^{\infty}\left(\rho_i^{-1}S_{i+1}+S_iC+\rho_k^{-1}S_i^{(k)}S_{k+1}\right)dt, \qquad k+1\leq i\leq n-1,\\ S_n(x)&=\int_x^{\infty}\left(S_nC+\rho_k^{-1}S_n^{(k)}S_{k+1}\right)dt+\left(\int_x^{\infty}\hat{q}\,dt\right)\hat{e}. \end{aligned}$$

We extend the definition of $S_{im}(x)$ to all of (a,∞) by setting $S_{im}(x)\equiv 0$ for $x>r_m$, $k+1\leq i\leq n$. We claim next that for each $i=k+1,\cdots,n$ the sequence $\{S_{im}(x)\}_{m=1}^{\infty}$ is uniformly bounded on compact subintervals of (a,∞). To verify this, we wish to examine the equations for the components of the S_{im}, which may be written as

$$(3.18)\qquad \begin{aligned} S_{im}^{(l)}(x)&=\int_x^{r_m}\left(\rho_i^{-1}S_{i+1,m}^{(l)}+\rho_{l-1}^{-1}S_{im}^{(l-1)}+\rho_{k-1}^{(k)}S_{im}^{(k)}S_{k+1,m}^{(l)}\right)dt,\\ S_{nm}^{(l)}(x)&=\int_x^{r_m}\left(\rho_{l-1}^{-1}S_{nm}^{(l-1)}+\rho_k^{-1}S_n^{(k)}S_{k+1,m}^{(l)}\right)dt+\begin{cases}\displaystyle\int_x^{r_m}q\,dt, & l=1,\\ 0, & l>1,\end{cases} \end{aligned}$$

for $l=1,\cdots,k$, $i=k+1,\cdots,n-1$.

For $i=k+1$ and $l=k$ we have

$$S_{k+1,m}^{(k)}(x)=g_m(x)+\int_x^{r_m}\rho_{k-1}^{-1}\left(S_{k+1,m}^{(k)}\right)^2dt$$

where

$$g_m(x)=\int_x^{r_m}\left(\rho_{k+1}^{-1}S_{k+2,m}^{(k)}+\rho_{k-1}^{-1}S_{k+1,m}^{(k-1)}\right)dt$$

is nonnegative and nonincreasing on $(a,r_m]$ with $g_m(r_m)=0$. Therefore, on any compact subinterval, since ρ_k^{-1} is bounded below, it follows from Lemma 3.2 that the sequence $\{S_{k+1,m}^{(k)}\}_{m=1}^{\infty}$ is uniformly bounded. Therefore, the sequence $\{g_m(x)\}_{m=1}^{\infty}$ must also be uniformly bounded on any compact subinterval of (a,∞). Thus, since the components of S_{im} are all nonnegative, it follows that the sequences $\{S_{k+2,m}^{(k)}\}_{m=1}^{\infty}$ and $\{S_{k+1,m}^{(k-1)}\}_{m=1}^{\infty}$ are also uniformly bounded on each compact subinterval. If we next examine the equation (3.18) for $l=k-1$ and $i=k+1$, we may conclude that the sequences $\{S_{k+2,m}^{(k-1)}\}_{m=1}^{\infty}$ and $\{S_{k+1,m}^{(k-2)}\}_{m=1}^{\infty}$ are also uniformly bounded on compact subintervals, and proceeding in this manner, we see that all of the sequences $\{S_{k+2,m}^{(j)}\}_{m=1}^{\infty}$ and $\{S_{k+1,m}^{(j)}\}_{m=1}^{\infty}$ for $1\le j\le k$ are uniformly bounded on compact subintervals. Similarly, we may now show that the sequences $\{S_{k+3,m}^{(j)}\}_{m=1}^{\infty},\cdots,\{S_{nm}^{(j)}\}_{m=1}^{\infty}$ are uniformly bounded on compact subintervals, for $1\le j\le k$. Thus, for each i, $\{S_{im}(x)\}_{m=1}^{\infty}$ is uniformly bounded on each compact subinterval. Moreover, it is easy to see that for each i, the sequence $\{S_{im}(x)\}_{m=1}^{\infty}$ is also equicontinuous on compact subintervals so that by the usual diagonalization procedure, we may select a subsequence, which we again label $\{S_{im}(x)\}$, which converges uniformly to $S_i(x)$, $i=k+1,\cdots,n$, on compact subintervals of (a,∞). It follows that the $S_i(x)$, $k+1\le i\le n$, solve system (3.17) on (a,∞).

Next choose a sequence $s_m\to\infty$ such that $Q(x)\ge Q(s_m)$ for $a<x\le s_m$. Then

$$\left|\int_x^{s_m}\frac{p}{\rho_0\rho_n}dt\right|=|P(x)-P(s_m)|\le|P(x)|\le|Q(x)|,\qquad a<x\le s_m,$$

where

$$P(x)\equiv\int_x^{\infty}\hat{p}\,dt,\qquad Q(x)\equiv\int_x^{\infty}\hat{q}\,dt.$$

Now consider the system

$$(3.19)_m\qquad\begin{aligned}\sigma_{im}(x)&=\int_x^{s_m}\left(\rho_i^{-1}\sigma_{i+1,m}+\sigma_{im}C+\rho_k^{-1}\sigma_{im}^{(k)}\sigma_{k+1,m}\right)dt,\qquad k+1\le i\le n-1,\\ \sigma_{nm}(x)&=\int_x^{s_m}\left(\sigma_{nm}C+\rho_k^{-1}\sigma_{nm}^{(k)}\sigma_{k+1,m}\right)dt+\left(\int_x^{s_m}\hat{p}\,dt\right)\hat{e}\end{aligned}$$

which we may write as

$$(3.20)_m\qquad\hat{\sigma}_m=T_m\hat{\sigma}_m$$

where $\hat{\sigma}_m$ is the $(n-k)\times k$ matrix

$$(3.21)\qquad\hat{\sigma}_m=\begin{pmatrix}\sigma_{k+1,m}\\ \sigma_{k+2,m}\\ \vdots\\ \sigma_{nm}\end{pmatrix}.$$

Again, as in Theorem 2.1, we notice that T_m is a self-map on $\mathfrak{N}_m$, the set of continuous $(n-k)\times k$ matrix-valued functions $\hat{\sigma}$ on $(a,s_m]$ with

$$|\sigma_i^{(j)}(x)|\le S_i^{(j)}(x),\qquad a<x\le s_m,\quad i=k+1,\cdots,n,$$

$j=1,\cdots,k$. The set $\mathfrak{D}_m$ is a closed convex subset of $\mathcal{C}_m$, the Fréchet space of continuous $(n-k)\times k$ matrix-valued functions on $(a,s_m]$ with the compact-open topology. The remainder of the argument is as in the proof of Theorem 2.1. If follows that (1.2) is $(k,n-k)$ disfocal on (a,s_m), $m=1,2,\cdots$. Therefore, since $(-1)^{n-k}p(x)<0$ on (a,∞), it follows that (1.2) is $(k,n-k)$ disfocal on (a,∞). This completes the proof of the theorem. □

REFERENCES

[1] S. AHMAD AND A. C. LAZER, *On an extension of the Sturm comparison theorem*, this Journal, 12 (1981), pp. 1–9.

[2] G. J. BUTLER, *Hille–Wintner type comparison theorems for 2nd order ordinary differential equations*, Proc. Amer. Math. Soc., 76 (1979), pp. 51–59.

[3] U. ELIAS, *Eigenvalue problems for the equation $Ly+p(x)y=0$*, J. Differential Equations, 29 (1978), pp. 28–57.

[4] ———, *Oscillatory solutions and extremal points for a linear differential equation*, Arch. Rational Mech. Anal., 70 (1979), pp. 177–198.

[5] ———, *Necessary conditions and sufficient conditions for disfocality and disconjugacy of a differential equation*, Pacific J. Math., 81 (1979), pp. 379–397.

[6] L. H. ERBE, *Hille–Wintner type comparison theorem for self-adjoint fourth order linear differential equations*, Proc. Amer. Math. Soc., 80 (1980), pp. 417–421.

[7] E. HILLE, *Non-oscillation theorems*, Trans. Amer. Math. Soc., 64 (1948), pp. 234–252.

[8] Z. NEHARI, *Nonlinear techniques for linear oscillation problems*, Trans. Amer. Math. Soc., 210 (1975), pp. 387–406.

[9] ———, *Green's functions and disconjugacy*, Arch. Rational Mech. Anal., 62 (1976), pp. 53–76.

[10] W. F. TRENCH, *Canonical forms and principle systems for general disconjugate equations*, Trans. Amer. Math. Soc., 189 (1974), pp. 319–327.

[11] A. WINTNER, *On the comparison theorem of Kneser–Hille*, Math. Scand., 5 (1957), pp. 255–260.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0006 $01.25/0

ON AN OSCILLATION THEOREM OF BELOHOREC*

MAN KAM KWONG† AND JAMES S. W. WONG‡

Abstract. An oscillation criterion is given for the second-order nonlinear differential equation $y''+a(t)|y|^{\gamma}\operatorname{sgn} y=0$, $0<\gamma<1$, where $a(t)$ is continuous but is not assumed to be nonnegative for all large values of t. This is an extension of a well-known result of Belohorec.

Key words. second order, nonlinear, differential equations, oscillation

AMS-MOS subject classification (1980). Primary 34C10, 34C15

Consider the second-order nonlinear differential equation

$$y''+a(t)|y|^{\gamma}\operatorname{sgn} y=0, \qquad \gamma\neq 1, \tag{1}$$

where $a(t)\in C[0,\infty)$. We restrict our attention to solutions of (1) which exist on some ray $[t_0,\infty)$, where $t_0\geq 0$ may depend on the particular solution. Such a solution is said to be *oscillatory* if it has arbitrarily large zeros. Equation (1) is called *oscillatory* if all continuable solutions are oscillatory. For a general discussion of nonlinear oscillation problems of type (1), we refer the reader to [10]. We are here concerned with sufficient conditions on $a(t)$ for (1) to be oscillatory when $a(t)$ is allowed to assume negative values for arbitrarily large values of t. The well-known Wintner oscillation criterion for the linear equation (i.e., (1)) with $\gamma=1$, states that if $a(t)$ satisfies

$$\lim_{T\to\infty}\int_0^T a(t)\,dt=+\infty, \tag{2}$$

then (1) is oscillatory for $\gamma=1$, see [8]. Customarily, (1) is called *superlinear* if $\gamma>1$ and *sublinear* when $0<\gamma<1$.

When $a(t)$ is nonnegative, stronger oscillation results exist for the nonlinear equation (1) when $\gamma\neq 1$, notably the following:

THEOREM A (Atkinson [1]). *Let* $\gamma>1$. *Then* (1) *is oscillatory if and only if*

$$\lim_{T\to\infty}\int_0^T ta(t)\,dt=+\infty. \tag{3}$$

THEOREM B (Belohorec [2]). *Let* $0<\gamma<1$. *Then* (1) *is oscillatory if and only if*

$$\lim_{T\to\infty}\int_0^T t^{\gamma}a(t)\,dt=+\infty. \tag{4}$$

When $a(t)$ is not assumed to be nonnegative such necessary and sufficient conditions need not hold. In fact, if $a(t)$ becomes negative on an open interval, then the nonlinear equation (1) always has noncontinuable solutions, when $\gamma>1$ (see Kiguradze [7]). However, Kiguradze [7] proved that condition (3) is sufficient that all continuable solutions of the superlinear equation are oscillatory. A similar result was established by Belohorec [3] for the sublinear equation, i.e., (1) with $0<\gamma<1$, namely that condition (4) is an oscillation criterion for all continuable solutions.

* Received by the editors June 19, 1981, and in revised form February 6, 1982.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115. The research of this author was partially supported by a grant from the Graduate School of Northern Illinois University.

‡ China Dyeing Works, Ltd., 833 Swire House, and Department of Mathematics, University of Hong Kong, Hong Kong.

The result of Kiguradze [7] was somewhat more general and can be stated as follows:

THEOREM C (Kiguradze [7]). *If there exists a positive function* $\varphi(t)$ *such that* $\varphi' \geq 0$ *and* $\varphi'' \leq 0$ *and satisfies*

$$\lim_{T\to\infty} \int_0^T \varphi(t)a(t)\,dt = +\infty, \tag{5}$$

and $\gamma > 1$, *then* (1) *is oscillatory, i.e. all continuable solutions are oscillatory.*

Attempts to extend Belohorec's result [3] can be found in Coles [5], who showed among other things that the condition

$$\lim_{T\to\infty} \int_0^T (t+k)^{\alpha} a(t)\,dt = +\infty, \tag{6}$$

where k is a constant and $0 \leq \alpha \leq \gamma$, is sufficient for the oscillation of (1) when $0 < \gamma < 1$. In fact, Belohorec [3] had proved that condition (6) with $k=0$ is sufficient for oscillation. The purpose of this note is to extend condition (6) in a manner analogous to that of Kiguradze's Theorem C. Our main result is

THEOREM 1. *If there exists a positive function* $\varphi(t)$ *such that* $\varphi' \geq 0$ *and* $\varphi'' \leq 0$ *and* $a(t)$ *satisfies*

$$\lim_{T\to\infty} \int_0^T \varphi^{\gamma}(t)a(t)\,dt = +\infty, \tag{7}$$

and $0 < \gamma < 1$, *then* (1) *is oscillatory.*

Proof. Assume the contrary. Then there exists a solution $y(t)$ which may be assumed to be positive on $[t_0, \infty)$ for some $t_0 \geq 0$. For $t \geq t_0$, define $z(t) = [y(t)/\varphi(t)]^{\gamma}$, which is again positive. Let $\beta = 1/\gamma > 1$; then $y(t) = \varphi(t)z^{\beta}(t)$. By simple differentiation, it is easy to verify that

$$\frac{1}{z}(\varphi z^{\beta})'' = \frac{\beta}{\beta-1}(\varphi z^{\beta-1})'' + \beta\varphi z^{\beta-3}z'^2 + \frac{1}{1-\beta}\varphi'' z^{\beta-1}, \tag{8}$$

which is the crucial step in this proof. Note that from (1) and the definition of $z(t)$, we have

$$\frac{y''}{z} = \frac{(\varphi z^{\beta})''}{z} = -a\varphi^{\gamma}. \tag{9}$$

Since $\varphi'' \leq 0$, $\beta > 1$, the last two terms in (8) are nonnegative; hence we may combine (8) and (9) and obtain the following inequality

$$\frac{\beta}{\beta-1}(\varphi z^{\beta-1})'' \leq -\varphi^{\gamma}a. \tag{10}$$

Integrating (10) twice from t_0 to t, we obtain

$$\varphi z^{\beta-1}(t) \leq C_1 + C_0 t - \frac{\beta-1}{\beta}\int_{t_0}^{t}\int_{t_0}^{s} \varphi^{\gamma}(\tau)a(\tau)\,d\tau\,ds, \tag{11}$$

where C_0, C_1 are appropriate integration constants. Clearly (7) implies that the right-hand side of (11) becomes eventually negative contradicting the assumption that $\varphi z^{\beta-1}$ is positive. This completes the proof.

Remark 1. We note that the above proof gives a stronger result. In fact, condition (7) can be weakened to that of

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T \int_0^t \varphi^{\gamma}(s)a(s)\,ds\,dt = +\infty, \tag{12}$$

which is sufficient to produce the desired contradiction in (11). In this regard, we generalize a result of Kamenev [6] for the sublinear equation by taking $\varphi(t)\equiv 1$. Similarly, Wintner's original result was proved under the following condition which is weaker than (2):

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T\int_0^t a(s)\,ds\,dt=+\infty. \tag{13}$$

Condition (13) has been shown by Butler [4] to be sufficient for oscillation of (1) in the superlinear case, $\gamma>1$. This is a major contribution in second-order nonlinear oscillation, see also [9]. In view of Butler's result and conditions (5) and (12), it would be tempting to conjecture that condition (13) may be weakened to

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T\int_0^t \varphi(s)a(s)\,ds\,dt=+\infty,$$

where $\varphi>0$, $\varphi'\geq 0$, $\varphi''\leq 0$ for the superlinear equation.

Remark 2. Consider (1) with $0<\gamma<1$ and $a(t)=t^{\lambda}\sin t$ or $t^{\lambda}\cos t$. It was shown by Butler [4] that (1) is oscillatory if $\lambda\geq 1$. Condition (12) with $\varphi(t)=t$ shows that (1) is also oscillatory if $1-\gamma<\lambda<1$. Butler conjectured that in this case (1) is oscillatory if and only if $\lambda\geq -\gamma$; see [4, p. 199] for further details.

REFERENCES

[1] F. V. ATKINSON, *On second order nonlinear oscillation*, Pacific J. Math., 5 (1955), pp. 643–647. MR 17 #264.

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

[3] _____, *Two remarks on the properties of solutions of a nonlinear differential equation*, Acta Fac. Rerum. Natur. Univ. Comen. Mathematica, XXII (1969), pp. 19–26.

[4] G. J. BUTLER, *Integral averages and the oscillation of second order ordinary differential equations*, this Journal, 11 (1980), pp. 190–200.

[5] W. J. COLES, *A nonlinear oscillation theorem*, International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 193–202.

[6] I. V. KAMENEV, *Certain specifically nonlinear oscillation theorems*, Mat. Zametki, 10 (1971), pp. 129–134. (In Russian).

[7] I. T. KIGURADZE, *A note on the oscillation of solutions of the equation* $u''+a(t)|u|^n \operatorname{sgn} u=0$, Casopis Pest Mat., 92 (1967), pp. 343–350. (In Russian).

[8] A. WINTNER, *A criterion of oscillatory stability*, Quart. Appl. Math., 7 (1949), pp. 115–117. MR 10 #456.

[9] J. S. W. WONG, *A second order nonlinear oscillation theorem*, Proc. Amer. Math. Soc., 40 (1973), pp. 487–491. MR 47 #7132.

[10] _____, *On the generalized Emden-Fowler equation*, SIAM Rev., 17 (1975), pp. 339–360. MR 51 #3610.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0007 $01.25/0

EXISTENCE AND UNIQUENESS OF SOLUTIONS OF SECOND ORDER NONLINEAR DIFFERENTIAL EQUATIONS*

JAMES T. SANDEFUR, JR.†

Abstract. A factoring technique is used to prove existence, uniqueness, and continuation properties for solutions to a class of second order semilinear differential equations in a Banach space. These results are then used to derive local and global existence results for a large class of partial differential equations. Among the examples considered are the semilinear versions of the wave equation (possibly damped or strongly damped), the telegraph equation, and the equation of motion for a vibrating plate.

Contrary to most techniques, this method does not require commutativity of the operators. An example of this is also given.

1. Introduction. Consider the abstract second order semilinear differential equation

$$u''(t)+Au'(t)+Bu(t)=f(t,u(t)) \tag{1.1}$$

in an arbitrary Banach space χ, where A and B are linear (in general unbounded) operators on χ. Typically A and B are differential operators, $\chi=L^p(\Omega)$, where Ω is some bounded or unbounded region in $\mathbf{R}^n$ (and usually $p=2$), and f is in some sense Lipschitz continuous, at least locally. Particular examples of (1.1) are semilinear versions of the wave equation (possibly damped or strongly damped), the telegraph equation, the vibrating beam equation, etc.

One standard approach to this type of problem is to reduce it to a first order system in some space $\chi_E\oplus\chi$, where $\chi_E\subset\chi$ has an "energy norm". The equation is then shown to be controlled by a local group giving existence and uniqueness of solutions on an interval $[-c,c]$, $c>0$. See for example [1], [4], [5], [7], [12]. One disadvantage to this approach is that the space χ_E depends on the particular equation. An advantage is that f need only be Lipschitz continuous with respect to the energy norm in u and may also be Lipschitz continuous in u'.

Some authors have approached (1.1) without reducing the order of the equation. For example Caughey and Ellison [2] use an eigenfunction expansion. Travis and Webb [11] use cosine functions in the case where $A=0$ and Lightbourne and Rankin [8] generalize this approach to cases where $A\neq 0$ using a combination of cosine functions and semigroups. These approaches have the advantage that they can be applied to large classes of equations. They also remain in the space χ.

Our approach is to factor equation (1.1) and then use semigroups and successive approximations to get existence and uniqueness. While this method may at first seem unnatural, its usefulness will be demonstrated by applying it to a large class of equations using only the quadratic equation and some freshman calculus. Moreover, the factoring procedure eliminates the need to find an energy norm suitable to the problem.

In §2 we will give preliminaries and our main result. In §3 we will apply our results to several examples while saving the proofs and further results for §4.

*Received by the editors March 10, 1982, and in revised form May 15, 1982.

†Center for Applied Mathematics, Cornell University, Ithaca, New York 14853 and Department of Mathematics, Georgetown University, Washington, DC 20057.

2. The main result. The set of bounded linear operators $\{T(t); t\in\mathbf{R}_+\}$, where $\mathbf{R}_+=[0,\infty)$, is a (C_0)-semigroup on χ if

(i) $T(t+s)=T(t)T(s)=T(s)T(t)$, $s,t\geq 0$.

(ii) $T(0)=I$ (the identity operator).

(iii) $T(\cdot)$ is strongly continuous in $t\in\mathbf{R}_+$.

(iv) $\|T(t)\|\leq Me^{\omega t}$ for some M, $\omega>0$, $t\in\mathbf{R}_+$.

The operator A is the generator of $T(\cdot)$ if $A\phi=\lim_{h\to 0^+}((T(h)-T(0))/h)\phi$ and $D(A)$, the domain of A, is the set of $\phi\in\chi$ for which the limit exists. Formally $T(t)\phi$ satisfies the Cauchy problem

$$\text{(CP)}\qquad u'(t)=Au(t),\qquad u(0)=\phi.$$

If $\phi\in D(A)$, then $u(\cdot)\in C^1(\mathbf{R}_+,\chi)$ and (CP) holds. More generally, $u(t)=T(t)\phi$ is said to be a *mild solution* of (CP) when $\phi\notin D(A)$.

Consider the Cauchy problem

$$\begin{aligned}&u''(t)-(A_1+A_2)u'(t)+A_2A_1u(t)=f(t,u(t)),\\ &u(0)=\phi,\qquad u'(0)=\psi,\end{aligned}\tag{2.1}$$

where A_1, A_2 are linear (possibly unbounded) operators on χ. A_1 and A_2 need not commute. Also assume that A_j generates the semigroup T_j, $j=1,2$. u is said to be a *mild solution* of (2.1) if it satisfies

$$\begin{aligned}u(t)=T_1(t)\phi+\int_0^t T_1(t-\tau)T_2(\tau)(\psi-A_1\phi)\,d\tau\\ +\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f(s,u(s))\,ds\,d\tau,\end{aligned}\tag{2.2}$$

where $\phi\in D(A_1)$. The idea for the integral equation (2.2) came from solving the Cauchy problem for

$$\frac{d}{dt}\begin{pmatrix}u_1(t)\\ u_2(t)\end{pmatrix}=\begin{pmatrix}A_1 & 1\\ 0 & A_2\end{pmatrix}\begin{pmatrix}u_1(t)\\ u_2(t)\end{pmatrix}+\begin{pmatrix}0\\ f(t,u_1(t))\end{pmatrix}\tag{2.3}$$

using the Phillips perturbation theorem. Equation (2.2) is the (mild) solution for the first component of (2.3) and first appeared in [10], with f dependent only on t.

It will be shown in §4, after some tedious calculations, that if $\phi\in D(A_jA_k)$, $\psi\in D(A_j)$, $j,k=1,2$ and $f(\cdot,u(\cdot))$ is twice continuously differentiable when u is, then u is a strong solution of (2.1).

Remark 2.1. In the case of the wave equation, i.e., $A_1=-A_2$, then $T_2(t)=T_1(-t)$ and we can define the cosine function $C(t)=(T(t)+T(-t))/2$. Then (2.2) simplifies to the semilinear wave equation of Travis and Webb [11].

Remark 2.2. Suppose that A_1 and A_2 also commute. The commuting means that $(\lambda_1 I-A_1)^{-1}$ and $(\lambda_2 I-A_2)^{-1}$ commute for all λ_j in the resolvent set of A_j, $j=1,2$. In this case we could also say that u is a mild solution of (2.1) if it satisfies

$$\begin{aligned}u(t)=\frac{1}{2}\Big[&(T_1(t)+T_2(t))\phi+\int_0^t T_1(t-\tau)T_2(\tau)(\psi-A_1\phi)\,d\tau\\ &+\int_0^t T_2(t-\tau)T_1(\tau)(\psi-A_2\phi)\,d\tau\\ &+\int_0^t\int_0^\tau\left[T_1(t-\tau)T_2(\tau-s)+T_2(t-\tau)T_1(\tau-s)\right]f(s,u(s))\,ds\,d\tau\Big],\end{aligned}\tag{2.4}$$

where $\phi \in D(A_1) \cap D(A_2)$. This is an average of (2.2) and (2.2) with the ones and twos switched. The advantage of (2.4) is the expected symmetry of the ones and twos.

HYPOTHESIS (H1). *f is a nonlinear mapping from $\mathbf{R}_+ \oplus \chi$ into χ. There is a positive nondecreasing function $g:[0,\infty)\to(0,\infty)$ such that $\|f(t,\phi)\| \le g(x)$, $\|f(t,\phi)-f(t,\psi)\| \le g(x)\|\phi-\psi\|$ if $\|\phi\|, \|\psi\| \le x$ and $t \in [0,T]$ for some $T>0$.*

THEOREM 2.1. *Suppose A_1 and A_2 are semigroup generators on χ, f satisfies* (H1), *$\phi \in D(A_1)$ and $\psi \in \chi$. Also assume that $f(\cdot, u(\cdot)) \in C(\mathbf{R}_+, \chi)$ when $u(\cdot) \in C(\mathbf{R}_+, \chi)$. Then there exists a unique continuous function u satisfying* (2.2) *on the interval $t \in [0,c]$ for some $c>0$.*

The trick is to use successive approximation. We define u_0 as the solution to the linearized equation,

$$u_0(t) = T_1(t)\phi + \int_0^t T_1(t-\tau)T_2(\tau)(\psi - A_1\phi)\,d\tau. \tag{2.5}$$

Then we define

$$u_{j+1}(t) = u_0(t) + \int_0^t \int_0^\tau S(t,s,\tau) f\big(s, u_j(s)\big)\,ds\,d\tau \tag{2.6}$$

where

$$S(t,s,\tau) = T_1(t-\tau)T_2(\tau-s). \tag{2.7}$$

We then show that u_j converges to a function u which satisfies (2.2). The entire proof will be given in §4 as well as global and asymptotic results. But first we will give some examples to illustrate the usefulness of this approach.

Remark 2.3. Note that when A_1 and A_2 generate groups we get existence on $[-c,c]$, $c>0$.

3. Examples. All of the examples will be done in the complex Hilbert space $L^2(\Omega)$, where Ω is either a smooth bounded region in $\mathbf{R}^n$ or all of $\mathbf{R}^n$, $n=1,2,3$. This will enable us to use the operational calculus and the associated spectral mapping theorem [3, p.1335] which we state here for convenience.

THEOREM 3.1. *Let A be a self-adjoint operator on a complex Hilbert space χ with spectrum $\sigma(A) \subseteq \mathbf{R}$ and let $\{E_\lambda : \lambda \in \mathbf{R}\}$ be its spectral resolution. Let g be a continuous function on $\sigma(A)$. Define $g(A)$ by $g(A)u = \int_{-\infty}^{\infty} g(\lambda)\,dE_\lambda u$ with domain $D(g(A)) = \{u \in \chi : \int_{-\infty}^{\infty} g(\lambda)\,dE_\lambda u$ exists$\}$. Then $g(A)$ is a closed, densely defined operator on χ and $\sigma(g(A)) = g(\sigma(A))$. If g_1 and g_2 are two such functions and $|g_1(\lambda)| \ge |g_2(\lambda)|$ for all $\lambda \in \sigma(A)$, then $D(g_1(A)) \subseteq D(g_2(A))$.*

It is an easy consequence of this that if $\mathrm{Re}\{\sigma(g(A))\}$ is bounded above, then $g(A)$ generates a (C_0)-semigroup of normal operators.

In the following examples, f is assumed to satisfy Hypothesis (H1).

Example 3.1. We start with a classic example, the semilinear wave equation:

$$\begin{aligned} &u_{tt}(t,x) - \Delta u(t,x) = f(t,u(t,x)), \\ &u(0,x) = \phi(x), \qquad u_t(0,x) = \psi(x), \end{aligned} \tag{3.1}$$

with $x \in \mathbf{R}^n$ and $\phi, \psi \in L^2(\mathbf{R}^n)$. Define $A = -\mathrm{cl}(\Delta)$, where cl means closure. It is well known that A is self-adjoint, and $\sigma(A) = \mathbf{R}_+$.

Let $g(\lambda)=i\sqrt{\lambda}$ and let $A_1=-A_2=g(A)$. A_1 and A_2 are skew adjoint and thus generate (C_0)-unitary groups T_j, where $T_j(t)=\exp(A_j t)$, $j=1,2, t\in\mathbf{R}$. By Theorem 2.1, u satisfies

$$u(t)=\int_0^\infty \exp\left(it\sqrt{\lambda}\right) d(E_\lambda\phi)+\int_0^t\int_0^\infty \exp\left(i(t-2\tau)\sqrt{\lambda}\right) d(E_\lambda\psi)\,d\tau$$
$$+\int_0^t\int_0^\infty -i\sqrt{\lambda}\exp\left(i(t-2\tau)\sqrt{\lambda}\right) d(E_\lambda\phi)\,d\tau$$
$$+\int_0^t\int_0^\tau\int_0^\infty \exp\left(i(t+s-2\tau)\sqrt{\lambda}\right) d(E_\lambda f(s,u(s)))\,ds\,d\tau.$$

By defining

$$C(t)\phi\equiv\int_0^\infty \cos\left(t\sqrt{\lambda}\right) d(E_\lambda\phi)=\frac{1}{2}\int_0^\infty \exp\left(it\sqrt{\lambda}\right)+\exp\left(-it\sqrt{\lambda}\right) d(E_\lambda\phi)$$

and

$$S(t)\phi\equiv\int_0^\infty \sin\left(t\sqrt{\lambda}\right) d(E_\lambda\phi),$$

we get that

$$u(t)=C(t)\phi+\int_0^t C(\tau)\psi\,d\tau+\int_0^t\int_0^\tau C(t-\tau)f(s,u(s))\,ds\,d\tau. \tag{3.2}$$

Note that the complex numbers have dropped out and that u is real. Also this solution exists on some interval $[-c,c]$, $c>0$, since T_j, $j=1,2$ are groups. For more details see [11].

Example 3.2. Consider (3.1) in $\chi=L^2(\Omega)$, where Ω is a bounded domain in $\mathbf{R}^n$ and its boundary $\partial\Omega$ is smooth, $n=1,2,3$. $A=-\mathrm{cl}(\Delta)$ restricted to $C_0^\infty(\Omega)$. Then A has a complete countable orthonormal set of eigenfunctions ϕ_1, $\phi_2,\cdots$ and eigenvalues λ_1, $\lambda_2,\cdots,\lambda_j>0$ and $\phi_j=0$ on $\partial\Omega$ for all j [3, p.1743]. Let the eigenfunction expansion for the initial values be $\phi=\sum_{j=1}^\infty a_j\phi_j$ and $\psi=\sum_{j=1}^\infty b_j\phi_j$. Since $S(t)\phi_j=\sin(t\sqrt{\lambda_j})\phi_j$ and $C(t)\phi_j=\cos(t\sqrt{\lambda_j})\phi_j$ by Example 3.1, we have the mild solution u satisfying

$$u(t)=\sum_{j=1}^\infty \Big[a_j\cos\left(t\sqrt{\lambda_j}\right)+b_j\sin\left(t\sqrt{\lambda_j}\right)/\sqrt{\lambda_j}$$
$$+\int_0^t \sin\left(t\sqrt{\lambda_j}-s\sqrt{\lambda_j}\right)\langle f(s,u(s)),\phi_j\rangle/\sqrt{\lambda_j}\,ds\Big]\phi_j,$$

where $\langle\psi,\phi\rangle=\int_\Omega\psi\phi^*\,dx$ and $*$ means complex conjugate. In the case where the eigenfunctions are known we can use successive approximation to estimate u.

We could also let $A=-\mathrm{cl}(\Delta)$ have zero Neumann conditions. The main difference is that $\lambda_1=0$.

Example 3.3. Consider the strongly damped semilinear wave equation

$$u_{tt}-\alpha\Delta u_t-\Delta u=f(t,u),\quad \alpha>0,\qquad u(0)=\phi,\quad u'(0)=\psi \tag{3.3}$$

in $L^2(\mathbf{R}^n)$ and again let $A=\mathrm{cl}(-\Delta)$. Define $g_\pm(\lambda)=(-\alpha\lambda\pm\sqrt{\alpha^2\lambda^2-4\lambda})/2$ and $A_\pm=g_\pm(A)$. It is easy to see that $\mathrm{Re}(g_\pm(\lambda))\le 0$ for $\lambda\in\sigma(A)=\mathbf{R}_+$ and therefore $\sigma(A_\pm)$ is in the left half plane. The semigroups generated by $A_\pm$ are $T_\pm(t)=\exp(tg_\pm(A))=\exp(tA_\pm)$. Equation (3.3) thus becomes

$$u_{tt}-(A_++A_-)u_t+A_+A_-u=f(t,u),\qquad u(0)=\phi,\quad u'(0)=\psi. \tag{3.4}$$

If $\{E_\lambda;\lambda\in\mathbf{R}\}$ is the spectral resolution for A, then

$$A_{\pm}\phi=\frac{1}{2}\int_0^{\beta}(-\alpha\lambda\pm i2h_+(\lambda))\,dE_\lambda\phi+\frac{1}{2}\int_{\beta}^{\infty}(-\alpha\lambda\pm 2h_-(\lambda))\,d(E_\lambda\phi),$$

where $\beta=4/\alpha^2$, $h_{\pm}(\lambda)=\sqrt{\pm(4\lambda-\alpha^2\lambda^2)}\,/2$ and the roots are real on the corresponding intervals. We can therefore combine integrals over $[0,\beta]$ into sines and cosines as in previous examples while leaving integrals over $[\beta,\infty)$ as exponentials. The final mild solution therefore satisfies

$$\begin{aligned}u(t)=&\int_0^{\beta}\exp(-\alpha\lambda t/2)\left[\cos(th_+(\lambda))+\alpha\lambda\sin(th_+(\lambda))/h_+(\lambda)\right]d(E_\lambda\phi)\\&+\int_0^{\beta}\exp(-\alpha t/2)\sin(th_+(\lambda))/h_+(\lambda)\,d(E_\lambda\psi)\\&+\int_0^t\int_0^{\beta}\exp(-\alpha\lambda(t-s)/2)\sin(h_+(\lambda)(t-s))/h_+(\lambda)d(E_\lambda f(s,u(s)))\,ds\\&+\int_{\beta}^{\infty}\exp(-\alpha\lambda t/2)\left[\cosh(th_-(\lambda))+\alpha\lambda\sinh(th_-(\lambda))/h_-(\lambda)\right]d(E_\lambda\phi)\\&+\int_{\beta}^{\infty}\exp(-\alpha\lambda t/2)\sinh(th_-(\lambda))/h_-(\lambda)\,d(E_\lambda\psi)\\&+\int_0^t\int_{\beta}^{\infty}\exp(-\alpha(t-s)/2)\sinh(h_-(\lambda)(t-s))/h_-(\lambda)\,d(E_\lambda f(s,u(s)))\,ds.\end{aligned}\tag{3.5}$$

Note that this solution exists only on $[0,c]$, $c>0$ since $A_{\pm}$ do not generate groups.

If we now consider (3.3) on $L^2(\Omega)$, where Ω is a smooth bounded region in $\mathbf{R}^n$, $n=1,2,3$, we have as in Example 3.2 an eigenfunction expansion for the solution. For additional results see [4], [9], [12].

Example 3.4. Consider the semilinear telegraph equation

$$u_{tt}+\alpha u_t-\Delta u=f(t,u)\qquad(\alpha>0)\tag{3.6}$$

in $L^2(\mathbf{R}^n)$ (or $L^2(\Omega)$, where Ω is smooth bounded in $\mathbf{R}^n$). Defining A as before, letting $g_{\pm}(\lambda)=(-\alpha\pm\sqrt{\alpha^2-4\lambda})/2$ and letting $A_{\pm}=g_{\pm}(A)$, we have that $\sigma(A_{\pm})\subset\{y:y=(-\alpha\pm\sqrt{\alpha^2-4\lambda})/2,\ \lambda\geq 0\}$ and $\operatorname{Re}\{\sigma(A_{\pm})\}\leq 0$. As in Example 3.3 we can let $\phi=\int_0^{\beta}\lambda\,d(E_\lambda\phi)$, where $\beta=\alpha^2/4$. Here the integrals over $[0,\beta]$ will be exponentials and the integrals over $[\beta,\infty)$ will contain sines and cosines. The solution will be the same as (3.5) except (i) the integrals $\int_0^{\beta}$ and $\int_{\beta}^{\infty}$ are interchanged, (ii) $(\alpha\lambda)$ becomes α and (iii) $h_{\pm}=\sqrt{\pm(4\lambda-\alpha^2)}\,/2$.

Example 3.5. The damped semilinear wave equation in one dimension is

$$\begin{aligned}&u_{tt}+\alpha u_{tx}-u_{xx}=f(t,u),\quad \alpha\in\mathbf{R},\quad x\in\mathbf{R},\\&u(0,x)=\phi(x),\quad u_t(0,x)=\psi(x),\quad \phi\text{ and }\psi\in L^2(\mathbf{R}).\end{aligned}\tag{3.7}$$

Here we let $A=\operatorname{cl}(\frac{1}{i}\frac{d}{dx})$ which is self-adjoint with $\sigma(A)=\mathbf{R}$. Let $A_{\pm}=g_{\pm}(A)$, where $g_{\pm}(\lambda)=(-\alpha\pm\sqrt{\alpha^2+4})i\lambda/2$. $\sigma(A_{\pm})$ is purely imaginary so that $A_{\pm}$ are skew adjoint and consequently generate unitary groups $T_{\pm}$ so that we get existence of a mild solution u on an interval $[-c,c]$, $c>0$.

Example 3.6. Consider Example 3.5 on $L^2([0,2\pi])$. Let $A=\frac{1}{i}\frac{d}{dx}$ with $D(A)=\{\phi;\phi\in L^2[0,2\pi],\ \phi$ is absolutely continuous and $\phi(0)=\phi(2\pi)\}$. Then A is self-adjoint, $\sigma(A)=\{j;j=1,\pm 1,\cdots\}$. An eigenvector ϕ_j associated with eigenvalue j is $\phi_j=e^{ijx}$.

Suppose we are given real initial data ϕ and ψ in terms of the orthonormal basis $\{\sin(nx)/\sqrt{\pi}, \cos(nx)/\sqrt{\pi}, 1/\sqrt{2\pi}\}$. In particular let $\phi=\cos(nx)$ and $\psi=0$. The general case is similar. Thus $\phi=(e^{inx}+e^{-inx})/2$ and u satisfies

$$\begin{aligned}u(t,x)&=\cos\left(tn\sqrt{\alpha^2+4}/2\right)\cos(nx-nt\alpha/2)-\alpha\sin\left(tn\sqrt{\alpha^2+4}/2\right)\sin(nx-tn\alpha/2)/\sqrt{4+\alpha^2}\\&\quad+\frac{2}{n\pi\sqrt{\alpha^2+4}}\int_0^t\sin\left((t-s)n\sqrt{\alpha^2+4}/2\right)\\&\qquad\cdot[\cos(nx+ns\alpha/2-nt\alpha/2)\langle f(s,u(s,y)),\cos(ny)\rangle\\&\qquad\quad+\sin(nx+ns\alpha/2-nt\alpha/2)\langle f(s,u(s,y)),\sin(ny)\rangle]\,ds,\end{aligned}$$

where $\langle f(s,u(s,y)),g(y)\rangle=\int_0^{2\pi}f(s,u(s,y))g^*(y)\,dy$.

The trick in this problem is that the calculations must be done using the eigenvectors $\{\exp(\pm inx)\}$ and then we convert back to the real basis $\{\cos(nx)\}$ and $\{\sin(nx)\}$, $n=1,2,\cdots$.

Example 3.7. Consider the equation of motion of a thin panel

$$\begin{aligned}&u_{tt}+\alpha u_{txxxx}+\sqrt{\rho}\,\delta u_t+u_{xxxx}-\Gamma u_{xx}+\rho u_x=f(t,u),\\&u(0,x)=\phi(x),\quad u_t(0,x)=\psi(x),\quad x\in[0,2\pi].\end{aligned}\tag{3.8}$$

Again let $A=\mathrm{cl}(\frac{1}{i}\frac{d}{dx})$ defined on $D(A)=\{\phi\in L^2[0,2\pi];\ \phi$ is absolutely continuous and $\phi(0)=\phi(2\pi)\}$. Then (3.8) becomes

$$u''+\left(\sqrt{\rho}\,\delta+A^4\right)u'+\left(A^4+\Gamma A^2+\rho iA\right)u=f(t,u).\tag{3.9}$$

Let

$$g_\pm(\lambda)=\frac{-\left(\sqrt{\rho}\,\delta+\lambda^4\right)\pm\sqrt{\left(\sqrt{\rho}\,\delta+\lambda^4\right)^2-4\left(\lambda^4+\Gamma\lambda^2+\rho i\lambda\right)}}{2}.$$

Again, with a bit of algebra, it can be shown that $\mathrm{Re}\{g_\pm\}$ are bounded above on $\sigma(A)=\mathbf{R}$, giving us existence of a mild solution on $[0,c]$ for some $c>0$. Also see [1], [7].

Example 3.8. We now consider a case in which the operators do not commute. Consider the Cauchy problem for

$$u_{tt}(t,x)+2b(x)u_t(t,x)-u_{xx}(t,x)=f(t,u(t))\tag{3.10}$$

in $\chi=L^2(\mathbf{R})$ with $b,b_x\in L^\infty(\mathbf{R})$. $A=\mathrm{cl}(\frac{d}{dx})$ is skew adjoint, and $B=b(x)$ and $B'=b'(x)$ are bounded linear operators on χ. Let $F(t,u(t))=f(t,u(t))+(B^2-B')u(t)$. F satisfies (H1) if f does.

Since $A_\pm=\pm A-B$ are bounded perturbations of a skew adjoint operator, they generate groups $T_\pm$. Note that $A_\pm$ do not commute since A and B do not commute. But by Theorem 2.1, we have existence and uniqueness of a mild solution to

$$\left(\frac{d}{dt}-A_+\right)\left(\frac{d}{dt}-A_-\right)u=F(t,u).\tag{3.11}$$

But (3.11) is equivalent to (3.10).

Note that the preceding can also be done in $L^2([0,1])$ with $D(A)=\{\phi\in L^2([0,1]);\ \phi$ is absolutely continuous and $\phi(0)=\phi(1)\}$.

4. Proofs and further results.

Proof of Theorem 2.1. Define u_j and $S(t,s,\tau)$ as in (2.5)–(2.7). For $0\le t\le T_0$, $\|T_j(t)\|\le M$ for $j=1,2$, some $M>0$ and $T_0=T$ (T as in Hypothesis H1). Let $k>2M\|\phi\|$. Then

$$\|u_0(t)\|\le M\|\phi\|+M^2\int_0^t\|\psi-A_1\phi\|\,d\tau\le M\|\phi\|+M^2t\|\psi-A_1\phi\|\le\frac{k}{2}$$

for $0\le t\le c(k)=c$ (and $0<c\le T_0$). Let $K=\sqrt{k/g(k)}\,/M$. If $K\ge c$ redefine g so that (H1) still holds and $K\le c$.

Assume that $\|u_j(t)\|\le k$ for $0\le t\le K$ and $j=0,\cdots,n$. Then

$$\begin{aligned}\|u_{n+1}(t)\|&\le\|u_0(t)\|+\int_0^t\int_0^\tau\|S(t,s,\tau)f(s,u_n(s))\|\,ds\,d\tau\\&\le k/2+\int_0^t\int_0^\tau M^2\|f(s,u_n(s))\|\,ds\,d\tau\le k/2+\int_0^t\int_0^\tau M^2g(k)\,ds\,d\tau\\&=k/2+M^2t^2g(k)/2\le k\end{aligned}$$

for $0\le t\le K$.Therefore $\|u_n(t)\|\le k$, for all n and $t\in[0,K]$,

$$\begin{aligned}\|u_2(t)-u_1(t)\|&\le M^2\int_0^t\int_0^\tau\|f(s,u_1(s))-f(s,u_2(s))\|\,ds\,d\tau\\&\le M^2\int_0^t\int_0^\tau g(k)\|u_1(s)-u_0(s)\|\,ds\,d\tau\\&\le M^2\int_0^t\int_0^\tau g(k)2k\,ds\,d\tau\\&=M^2t^2g(k)k\end{aligned}$$

by (H1), the above and the fact that $\|S(t,s,\tau)\|\le M^2$ on $[0,K]$. Assume for $j=1,\cdots,n-1$ and $t\in[0,K]$ that

$$\|u_{j+1}(t)-u_j(t)\|\le\bigl(g(k)M^2t^2\bigr)^j2k/(2j)!.$$

Then

$$\begin{aligned}\|u_{n+1}(t)-u_n(t)\|&\le M^2\int_0^t\int_0^\tau\|f(s,u_n(s))-f(s,u_{n-1}(s))\|\,ds\,d\tau\\&\le M^2\int_0^t\int_0^\tau g(k)\|u_n(s)-u_{n-1}(s)\|\,ds\,d\tau\\&\le\int_0^t\int_0^\tau M^{2n}g^n(k)s^{2n-2}2k/(2n-2)!\,ds\,d\tau\\&=g^n(k)(Mt)^{2n}2k/(2n)!\end{aligned}$$

for $t\in[0,K]$. Hence $\{u_j\}$ is strongly convergent to a function u uniformly on $t\in[0,K]$. By letting $j\to\infty$ in (2.6) we see that u satisfies (2.2) on $[0,K]$.

If $v(t)$ is another solution of (2.2) and $\|v(t)\|\le k'$ on $t\in[0,K']$, $K'\le T$ (T as in hypothesis H1), then

$$\begin{aligned}\|u(t)-v(t)\|&\le M^2\int_0^t\int_0^\tau\|f(s,u(s))-f(s,v(s))\|\,ds\,d\tau\\&\le M^2g(k'')\int_0^t\int_0^\tau\|u(s)-v(s)\|\,ds\,d\tau,\end{aligned}$$

where $k''=\max\{k,k'\}$. But the right side equals

$$M^2g(k'')\int_0^t\int_s^t\|u(s)-v(s)\|\,d\tau\,ds=M^2g(k'')\int_0^t(t-s)\|u(s)-v(s)\|\,ds$$

$$\leq M^2g(k'')t\int_0^t\|u(s)-v(s)\|\,ds$$

on $t\in[0,K'']$, where $K''=\min\{K,K'\}$. From this it follows that $u(t)=v(t)$ on their common domain. The continuity of u is straight-forward and will be omitted. Q.E.D.

An equivalent proof of this theorem would be to use the Picard-Banach fixed point theorem on (2.2).

We now discuss strong solutions of (2.1). We say that a function $u:\mathbf{R}_+\to\chi$ is a strong solution of (2.1) if $u\in C^1(\mathbf{R}_+,\chi)\cap C(\mathbf{R}_+,D(A_1))$, $(u'-A_1u)\in C^1(\mathbf{R}_+,\chi)\cap C(\mathbf{R}_+,D(A_2))$ and u satisfies $((d/dt)-A_2)((d/dt)-A_1)u(t)=f(t,u(t))$.

THEOREM 4.1. *Suppose* $u(t)\in C^2(\mathbf{R}_+,\chi)$ *and that* u *satisfies* (2.2), *where* $\phi\in D(A_jA_k)$, $\psi\in D(A_j)$, $j,k=1,2$. *Also assume that* $f(\cdot)=f(\cdot,v(\cdot))\in C^2(\mathbf{R}_+,\chi)$ *whenever* $v\in C^2(\mathbf{R}_+,\chi)$. *Then* u *is a strong solution to* (2.1).

Proof. Let $f_0(t)=f(t,u(t))$ and $f_0'(t)=df_0(t)/dt$, $f_0''(t)=d^2f_0(t)/dt^2$. Also let $w(t)=\int_0^t T(t-s)g(s)\,ds$, where T is the semigroup generated by some operator A. We first give several formulas.

LEMMA 4.2. *If* $g\in C^1(\mathbf{R}_+,\chi)$, *then* $w\in C(\mathbf{R}_+,D(A))\cap C^1(\mathbf{R}_+,\chi)$. *Also*

$$w'(t)=T(t)g(0)+\int_0^t T(t-s)g'(s)\,ds \tag{4.1}$$

and

$$Aw(t)=T(t)g(0)-g(0)+\int_0^t[-g'(s)+T(t-s)g'(s)]\,ds. \tag{4.2}$$

For the proof see Goldstein [6, p. 49].

Now let

$$v(t)=\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f_0(s)\,ds\,d\tau$$

$$=\int_0^t T_1(t-\tau)\int_0^\tau T_2(\tau-s)f_0(s)\,ds\,d\tau.$$

Let $g(\tau)=\int_0^\tau T_2(\tau-s)f(s,u(s))\,ds$. By applying (4.1) to v and then to g when computing g', we get

$$v'(t)=\int_0^t T_1(t-\tau)T_2(\tau)f_0(0)\,d\tau+\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f_0'(s)\,ds\,d\tau. \tag{4.3}$$

Applying (4.2) then (4.1) to v gives

$$\begin{aligned}A_1v(t)=&-\int_0^t T_2(\tau)f_0(0)\,d\tau-\int_0^t\int_0^\tau T_2(\tau-s)f_0'(s)\,ds\,d\tau\\&+\int_0^t T_1(t-\tau)T_2(\tau)f_0(0)\,d\tau+\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f_0'(s)\,ds\,d\tau.\end{aligned} \tag{4.4}$$

Therefore

$$v'(t)-A_1v(t)=\int_0^t T_2(\tau)f_0(0)\,d\tau+\int_0^t\int_0^\tau T_2(\tau-s)f_0'(s)\,ds\,d\tau. \tag{4.5}$$

Repeating this process gives

$$\frac{d}{dt}(v'(t)-A_1v(t))=T_2(t)f_0(0)+\int_0^t T_2(\tau)f_0'(0)\,d\tau + \int_0^t\int_0^\tau T_2(\tau-s)f_0''(s)\,ds\,d\tau. \tag{4.6}$$

By similar arguments we can show that

$$A_2(-v'(t)+A_1v(t))=-T_2(t)f_0(0)-\int_0^t T_2(\tau)f_0'(0)\,d\tau - \int_0^t\int_0^\tau T_2(\tau-s)f_0''(s)\,ds\,d\tau+f(t). \tag{4.7}$$

Adding (4.6) and (4.7) gives

$$\left(\frac{d}{dt}-A_2\right)\left(\frac{d}{dt}-A_1\right)v(t)=f(t). \tag{4.8}$$

Let

$$\tilde{u}(t)=T_1(t)\phi+\int_0^t T_1(t-\tau)T_2(\tau)(\psi-A_1\phi)\,d\tau.$$

Then

$$\left(\frac{d}{dt}-A_2\right)\left(\frac{d}{dt}-A_1\right)\tilde{u}(t)=0.$$

Therefore $u=\tilde{u}+v$ is a strong solution to (2.1). Q.E.D.

Remark 4.1. If A_1 and A_2 commute and $f_0(0)\in D(A_1)\cap D(A_2)$ in addition to the conditions of Theorem 4.1, then $u\in C^2(\mathbf{R}_+,\chi)\cap C(\mathbf{R}_+,D(A_1A_2))$, $u'\in C(\mathbf{R}_+,D(A_1)\cap D(A_2))$ and u satisfies

$$u''(t)-(A_1+A_2)u''(t)+A_1A_2u(t)=f(t,u(t)).$$

Now we discuss global solutions.

HYPOTHESIS (H2). *f is a nonlinear jointly continuous mapping from $R_+\times\chi$ into χ. There is a positive nondecreasing function $g:[0,\infty)\to(0,\infty)$ such that $\|f(t,\phi)-f(t,\psi)\|\le g(\tau)\|\phi-\psi\|$ for $0\le t\le\tau$.*

THEOREM 4.3. *Suppose A_1 and A_2 are semigroup generators on χ and f satisfies* (H2). *Also assume $\phi\in D(A_1)\cap D(A_2)$. Then there exists a unique continuous solution to* (2.2) *on $\mathbf{R}_+$.*

Proof. Define $u_j(t)$ again by (2.5) and (2.6). For $0\le t\le T$, $\|u_1(t)-u_0(t)\|\le M^2\int_0^t\int_0^\tau\|f(s,u_0(s))\|\,ds\,d\tau$. Since $\|f(s,u_0(s))-f(s,0)\|\le g(T)\|u_0(s)\|$ and $\|f(s,0)\|$ and $\|u_0(s)\|$ are bounded on $[0,T]$ we have $\|u_1(t)-u_0(t)\|\le M^2g(T)Ct^2/2$ for some $C>0$. Suppose $\|u_j(t)-u_{j-1}(t)\|\le(M^2g(T)t^2)^jC/(2j)!$ for $j=1,2,\cdots,n$. Then

$$\begin{aligned}\|u_{n+1}(t)-u_n(t)\|&\le M^2\int_0^t\int_0^\tau\|f(s,u_n(s))-f(s,u_{n-1}(s))\|\,ds\,d\tau\\&\le M^2g(T)\int_0^t\int_0^\tau\|u_n(s)-u_{n-1}(s)\|\,ds\,d\tau\\&\le M^2g(T)\int_0^t\int_0^\tau\left(M^2g(T)t^2\right)^n\frac{C}{(2n)!}\,ds\,d\tau\\&=\left(M^2g(T)t^2\right)^{n+1}\frac{C}{(2n+2)!}\end{aligned}$$

completing the induction. Therefore $\{u_j(t)\}$ is strongly convergent on $[0,T]$. Since T is arbitrary, $\{u_j(t)\}$ converges on $[0,\infty)$ to a solution u to (2.2), uniformly on compact subsets.

The uniqueness and continuity follow from Theorem 2.1. Q.E.D.

We next prove the continuation property.

THEOREM 4.4. *Suppose A_1 and A_2 are semigroup generators on χ, $f(\cdot,u(\cdot))\in C(\mathbf{R}_+,\chi)$ when $u(\cdot)\in C(\mathbf{R}_+,\chi)$ and $\phi\in D(A_1)\cap D(A_2)$. Also assume that for each $T>0$ there exists a positive nondecreasing function $g_T:[0,\infty)\to(0,\infty)$ such that* (H1) *is satisfied. Let $T_0>0$ be such that there exists a solution u to* (2.2) *on $[0,T_0)$ but that u cannot be continued beyond $[0,T_0]$. Then either $T_0=+\infty$ or $\lim_{t\to T_0^-}\sup\|u(t)\|=+\infty$.*

Proof. Let $T_0=T$, and suppose that $T<\infty$ and $\|u(t)\|\le M$ for $0\le t<T$. We will first show that $\operatorname{st}-\lim_{t\to T^-}u(t)$ $(\equiv u(T))$ exists. Let

$$v(t)=\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f(s,u(s))\,ds\,d\tau.$$

For $0\le t<\hat{t}<T$,

$$\begin{aligned}v(t)-v(\hat{t})&=\int_0^t\int_0^\tau T_1(t-\tau)\big(I-T_1(\hat{t}-t)\big)T_2(\tau-s)f(s,u(s))\,ds\,d\tau\\&\quad+\int_t^{\hat{t}}\int_0^\tau T_1(\hat{t}-\tau)T_2(\tau-s)f(s,u(s))\,ds\,d\tau\\&\equiv J_1+J_2,\end{aligned}$$

$$\|J_1\|\le\|I-T_1(\hat{t}-t)\|\left\|\int_0^t\int_0^\tau T_1(t-\tau)T_2(\tau-s)f(s,u(s))\,ds\,d\tau\right\|.$$

Since $\|T_1(\cdot)\|$ and $\|T_2(\cdot)\|$ are bounded on $[0,T]$,

$$\begin{aligned}\|J_1\|&\le C\|I-T_1(\hat{t}-t)\|\int_0^t\int_0^\tau g_T(M)\|u(s)\|\,ds\,d\tau\\&\le\frac{CMg_T(M)T^2}{2}\|I-T_1(\hat{t}-t)\|.\end{aligned}$$

By the continuity of T_1, $\|J_1\|\le\frac{\varepsilon}{2}$ for $|\hat{t}-t|<\delta$ for some $\delta>0$.

Likewise $\|J_2\|\le CMg_T(M)(\hat{t}^2-t^2)/2<\frac{\varepsilon}{2}$ for $|\hat{t}-t|<\delta_1$ for some $\delta_1>0$. Therefore $\|v(t)-v(\hat{t})\|<\varepsilon$ for $|t-\hat{t}|<\min\{\delta,\delta_1\}$ and $0\le t<\hat{t}<T$. Therefore $\lim_{t\to T^-}v(t)$ exists and consequently $\lim_{t\to T^-}u(t)$ exists.

Now we wish to extend u to the interval $[0,T+\varepsilon)$ for some $\varepsilon>0$. Define $u_j(t)=u(t)$ for $t\in[0,T]$ and $j=0,1,\cdots$. For $t>T$ define

$$\begin{aligned}u_0(t)&=T_1(t)\phi+\int_0^t T_1(t-\tau)T_2(\tau)(\psi-A_1\phi)\,d\tau\\&\quad+\int_0^T\int_0^\tau T_1(t-\tau)T_2(\tau)f(s,u(s))\,ds\,d\tau\\&\quad+\int_T^t\int_0^T S(t,s,\tau)f(s,u(s))\,ds\,d\tau.\end{aligned}$$

Define

$$u_{j+1}(t)=u_0(t)+\int_T^t\int_T^\tau S(t,s,\tau)f(s,u_j(s))\,ds\,d\tau.$$

In a manner similar to the proof of Theorem 2.1 we show that for some $t_0 > T$, $\{u_j(t)\}$ converges strongly to a function u uniformly on $t \in [0, t_0]$. By taking limits and observing that the integrals involving $f(s, u(s))$ add up to $\int_0^t \int_0^\tau S(t,s,\tau) f(s,u(s))\, ds\, d\tau$, we see that u satisfies (2.2) on $[0, t_0]$. Q.E.D.

Acknowledgment. I wish to thank Professor Jerome A. Goldstein for his helpful comments on this paper. I also wish to thank Cornell University and, in particular, Professors Lawrence Payne and Philip Holmes for their hospitality.

REFERENCES

[1] J. M. BALL, *Stability theory for an extensible beam*, J. Differential Equations, 14 (1973), pp. 399–418.
[2] T. K. CAUGHEY AND J. ELLISON, *Existence, uniqueness and stability of solutions of a class of nonlinear partial differential equations*, J. Math. Anal. Appl., 51 (1975), pp. 1–32.
[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part II, Interscience, New York, 1963.
[4] W. E. FITZGIBBON, *Strongly damped quasilinear evolution equations*, J. Math. Anal. Appl., 79 (1981), pp. 536–550.
[5] J. A. GOLDSTEIN, *Semigroups and second-order differential equations*, J. Funct. Anal., 4 (1969), pp. 50–70.
[6] _____, *Semigroups of operators and abstract Cauchy problems*, lecture notes, Tulane University, New Orleans, 1970.
[7] P. HOLMES AND J. MARSDEN, *Bifurcation to divergence and flutter in flow-induced oscillations: an infinite dimensional analysis*, Automatica, 14 (1978), pp. 367–384.
[8] J. H. LIGHTBOURNE, III AND S. M. RANKIN III, *Cosine families and damped second order differential equations*, to appear.
[9] P. MASSATT, *Limiting behavior for strongly damped nonlinear wave equations*, in press.
[10] J. T. SANDEFUR, *Higher order abstract Cauchy problems*, J. Math. Anal. Appl., 60 (1977), pp. 728–742.
[11] C. C. TRAVIS AND G. F. WEBB, *Cosine families and abstract nonlinear second order differential equations*, Acta Math. Acad. Sci. Hung., 32 (1978), pp. 75–96.
[12] G. F. WEBB, *Existence and asymptotic behavior for strongly damped nonlinear wave equations*, Canad. J. Math., 32 (1980), pp. 631–643.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0008 $01.25/0

THE CONNECTION BETWEEN PARTIAL DIFFERENTIAL EQUATIONS SOLUBLE BY INVERSE SCATTERING AND ORDINARY DIFFERENTIAL EQUATIONS OF PAINLEVÉ TYPE*

J. B. McLEOD† AND P. J. OLVER‡

Abstract. A completely integrable partial differential equation is one which has a Lax representation, or, more precisely, can be solved via a linear integral equation of Gel'fand–Levitan type, the classic example being the Korteweg–de Vries equation. An ordinary differential equation is of Painlevé type if the only singularities of its solutions in the complex plane are poles. It is shown that, under certain restrictions, if G is an analytic, regular symmetry group of a completely integrable partial differential equation, then the reduced ordinary differential equation for the G-invariant solutions is necessarily of Painlevé type. This gives a useful necessary condition for complete integrability, which is applied to investigate the integrability of certain generalizations of the Korteweg–de Vries equation, Klein–Gordon equations, some model nonlinear wave equations of Whitham and Benjamin, and the BBM equation.

Key words. Completely integrable partial differential equations, inverse scattering method, Gel'fand–Levitan equation, KdV equation, Klein–Gordon equations, nonlinear Schrödinger equation, similarity solutions, Painleve transcendents

1. Introduction. The recent discovery of nonlinear partial differential equations which can be exactly solved by the linear integral equations of inverse scattering theory has provoked considerable interest in the range of applicability of these methods for the integration of nonlinear equations in mathematical physics. The original investigations of Gardner, Kruskal and Miura [26] and Lax [22] for the Korteweg–de Vries (KdV) equation have now been extended to solve a surprising number of differential equations of physical interest, including the sine-Gordon, nonlinear Schrödinger, three-wave interaction and other equations (cf. [1], [19], [37], [38], [39]). In all of these examples, the given equation is recast into a "Lax representation,"

$$\frac{dL}{dt}=[B,L]=BL-LB, \tag{1.1}$$

where L, B are linear differential operators depending on the solution $u(x,t)$ of the equation, with B skew-adjoint. This representation implies that the spectrum of L has an elementary time evolution, and hence the original equation can be integrated once the inverse scattering problem of reconstructing the potential $u(x,t)$ from the spectral data of the corresponding operator L has been solved. In all known examples, this inverse scattering problem is effected through the solution of a linear integral equation of the form

$$K(x,y;t)+F(x,y;t)+\int_x^\infty K(x,z;t)H(z,y;t)\,dz=0, \tag{1.2}$$

known as the *Gel'fand–Levitan equation*. Here F and H are constructed from the spectral data of L; the potential $u(x,t)$ is recovered from the values of K on the diagonal $x=y$.

* Received by the editors November 5, 1980, and in revised form January 18, 1982. This research was sponsored by the United States Army under contract DAAG29-80-C-0041.

† Mathematical Institute, University of Oxford, Oxford, United Kingdom.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The research of this author was supported in part by the National Science Foundation under grant MCS 81-00786.

Hereafter, any partial differential equation which can be solved by such a linear integral equation will be termed *completely integrable*, this terminology stemming from the interpretation of the KdV equation as a completely integrable Hamiltonian system [11], [23]. Of course, only certain types of solutions can be obtained in this fashion, so this definition is subject to further refinement (cf. Definition 2.1). Completely integrable equations all seem to have many other remarkable properties in common including cleanly interacting soliton solutions, existence of infinitely many conservation laws, Bäcklund transformations, etc. (cf. [21]). However, the precise interrelationship among these properties remains to be rigorously formulated; thus reasons of practicality necessitate the adoption of the Gel'fand–Levitan type of linear integral equation as the distinguishing characteristic of complete integrability.

The most notable drawback in the applicability of inverse scattering techniques is that there is as yet no systematic method for determining whether a given differential equation is completely integrable, i.e., can be solved by such a linear integral equation. In this paper we find a useful necessary condition for integrability based on the nature of the complex singularities of group-invariant solutions to the equation. Whereas we are thus no closer to finding a scattering problem if it exists, this condition is useful for determining when no such solution is possible. In the applications to be considered, a number of nonlinear partial differential equations (p.d.e.'s) of interest will be shown not to be integrable by inverse scattering methods.

This condition was inspired by an observation of Ablowitz, Ramani and Segur [2], [4] that the ordinary differential equations for group-invariant (self-similar) solutions of known examples of completely integrable equations inevitably are equations of the type studied by Painlevé and his students; these are characterized by the property that all their solutions are meromorphic in the complex plane (cf. [17], [18]). Such an equation will be referred to as an equation of Painlevé type. (Painlevé also allowed fixed singularities of arbitrary type, but we will not.) This leads immediately to the conjecture proposed by Ablowitz, Ramani and Segur [2] and Hastings and McLeod [16]:

CONJECTURE. *If a system of partial differential equations is completely integrable, and G is a symmetry group of this system, then the reduced system of ordinary differential equations for the G-invariant solutions is of Painlevé type.*

This conjecture, if true, would provide a powerful necessary condition to test for complete integrability. Here we will prove a somewhat weakened version of the conjecture, which nevertheless proves useful in several applications. There are two restrictions. First, if, in the Lax operator L, some combination of the solution u and its spatial derivatives occurs, say $Q(u)$, then it is this combination (or combinations) that must have only poles as singularities. For instance, if $L=D^2+u_x$, then only u_x is required to have poles, and thus we may allow logarithmic branch points as singularities of the solutions of the reduced ordinary differential equations. Usually we will assume that Q is a linear combination of u and its spatial derivatives, calling this case *linearly completely integrable*. Secondly, the same combination Q must satisfy certain preconditions for the inverse scattering formalism to go through; this means that, when restricted to the real axis, Q either is periodic or satisfies decay conditions at $x=\pm\infty$, which implies corresponding restrictions on the solutions u that can be considered. It is only for such solutions such that $Q(u)$ must be meromorphic. If a system of ordinary differential equations has the property that, for such solutions u, the combination $Q(u)$ is meromorphic, we say that the system is of *restricted Painlevé type* relative to Q. Our basic result, in rough form, replaces "Painlevé type" by "restricted Painlevé type" in the above conjecture.

The main tool in our proof is a theorem of Steinberg [32] which states that if $T(z)$ is an analytic family of compact operators in a Banach space, then $(I-T(z))^{-1}$, provided this inverse exists for at least one value of z, is a meromorphic family of operators. Under appropriate assumptions on the initial data of our completely integrable system (to ensure that the functions F and H in the Gel'fand–Levitan equation satisfy certain analyticity criteria) we can conclude from Steinberg's result that Q must be a meromorphic function of (x,t). Now suppose that G is a one-parameter, analytic, regular local group of transformations acting on the space of independent and dependent variables which leaves the set of solutions of the system of partial differential equations invariant. Then the G-invariant (self-similar) solutions can all be found by integrating a system of ordinary differential equations on the quotient manifold whose points correspond to the orbits of G. The analyticity of G implies that for any G-invariant solution whose initial data satisfies the inverse scattering assumptions, the function Q on the quotient manifold can have only poles for singularities. In other words, the reduced system of ordinary differential equations must be of restricted Painlevé type relative to Q.

Ablowitz, Ramani and Segur [2], [3] have also given proofs of a version of the above conjecture. They restrict their attention to Gel'fand–Levitan equations of Fredholm type, and their groups are only groups of scaling transformations. Thus our result is somewhat more general. Both proofs are necessarily restricted to certain types of solutions, in particular, solutions decaying sufficiently rapidly as $|x|\to\infty$ are allowed. Extensions to the case of spatially periodic solutions can be inferred from the work of McKean and Trubowitz on the Korteweg–de Vries equation [23], [34], although the analogue of the Gel'fand–Levitan equation is not explicitly written down. We strongly suspect, however, that solutions are in general meromorphic in the periodic case also, and therefore include solutions of this type in our test for complete integrability. It would be of great interest to remove all restrictions on the types of solutions for which such a result can be proved and thereby prove the complete version of the conjecture.

In § 3 we discuss some applications of this result. First we show that the generalized KdV equation

$$u_t+u^p u_x+u_{xxx}=0 \tag{1.3}$$

can be linearly completely integrable only if $p=0,1$, or 2. These exceptional cases correspond to the Airy equation in moving coordinates, the KdV, and the modified KdV equations, which are well known to be completely integrable. Secondly we consider a nonlinear Klein–Gordon equation in characteristic coordinates:

$$u_{xt}=f'(u). \tag{1.4}$$

It is shown that if $f(u)$ is a rational function, real for real u and with two consecutive zeros, simple or double, on the real axis, and if (1.4) is linearly completely integrable, then f is a polynomial of degree at most 4. Also, if $f(u)$ is a linear combination of exponentials $e^{\alpha_j u}$ with the α_j all rational multiples of some complex number α, again real for real u and with two consecutive simple or double zeros, and if (1.4) is linearly completely integrable, then

$$f(u)=c_2e^{2\beta u}+c_1e^{\beta u}+c_0+c_{-1}e^{-\beta u}+c_{-2}e^{-2\beta u}$$

for some β. This includes the sine- and sinh-Gordon, and an equation due to Mikhailov [24], [25], which are known to be integrable, and the double sine-Gordon equation, whose status is a matter of dispute. The next application shows that certain nonlinear

model wave equations considered by Benjamin, Bona and Mahony [5] and Whitham [36] cannot be linearly completely integrable. The last example deals with the BBM equation [5],

$$u_t + uu_x - u_{xxt} = 0. \tag{1.5}$$

Although this cannot be treated rigorously by the methods of the present paper, we show that if the full conjecture were true, then (1.5) could not be linearly completely integrable.

Finally we discuss the general Lax representations of Gel'fand and Dikii for scalar differential operators L of order n (see [12], [13]). For n a composite number, there exist steady state solutions of the corresponding evolutionary systems with arbitrary complex singularities. This suggests that the inverse scattering problem for such an L is not amenable to solution by a Gel'fand–Levitan type equation, at least in the form discussed here. Indeed, only for second and third order L (see [20]) has the inverse problem been solved, so the theory for fourth order operators becomes of great interest. From those results, it can be seen that our criterion for complete integrability is a powerful preliminary test to determine whether a given system can be integrated by inverse scattering.

2. Analyticity properties of completely integrable differential equations. Consider a system of partial differential equations

$$\Delta(t,x,u) = 0, \tag{2.1}$$

where $x, t \in \mathbf{R}$ and $u = (u^1, \cdots, u^m) \in \mathbf{R}^m$ is a vector-valued function. We assume that the initial value problem of (2.1) with

$$u(x,0) = f(x) \tag{2.2}$$

is well posed for f in some Banach space $\mathfrak{B}$ of functions, so that for t sufficiently small, there is a unique solution $u(x,t)$ of (2.1)–(2.2). In practice $\mathfrak{B}$ is either a space of functions decreasing sufficiently rapidly at $\pm\infty$ or a space of periodic functions. Usually the presence of appropriate conservation laws will ensure that the solutions are actually global in t, but this will not be assumed a priori. The first task is to make precise what is meant by (2.1) being completely integrable.

DEFINITION 2.1. A system of partial differential equations is *completely integrable relative to* $Q(u)$ *in the Banach space* $\mathfrak{B}$ if there is a linear matrix integral equation of the form

$$K(x,y;t) + F(x,y;t) + \int_x^\infty K(x,z;t) H(z,y;t)\, dz = 0, \tag{2.3}$$

called the *Gel'fand–Levitan equation*, satisfying the following properties:

i) F, H, K are $N \times N$ matrices of functions;

ii) F and H are uniquely determined by the initial data (2.2);

iii) for initial data in $\mathfrak{B}$, and for all real x, y, all complex ε, and t in some domain Ω in $\mathbf{C}$, the functions $F(x-\varepsilon t, y-\varepsilon t; t)$ and $H(x-\varepsilon t, y-\varepsilon t; t)$ are analytic in ε, t, and there is a Banach space $\mathfrak{B}^*$ (not necessarily the same as $\mathfrak{B}$) for which $F(x-\varepsilon t, y-\varepsilon t; t) \in \mathfrak{B}^*$ as a function of y and the operator

$$T(x,t) f(y) = \int_x^\infty f(z) H(z-\varepsilon t, y-\varepsilon t; t)\, dz$$

is a compact operator in $\mathfrak{B}^*$;

iv) the Gel'fand–Levitan equation has a unique solution (in $\mathfrak{B}^*$) for all x and at least one t in Ω;

v) the solution u of the system (2.1), (2.2) can be recovered from the solution K of the Gel'fand–Levitan equation via a relation of the form

$$Q[u(x,t)]=P[K(x,x,t)], \tag{2.4}$$

where Q is some function of u and its spatial derivatives, and P is a polynomial in K and its spatial derivatives.

Thus to recover the solution u of a completely integrable system of partial differential equations, we must solve the Gel'fand–Levitan equation for K and then solve the differential equation (2.4) for u. In practical examples, Q is a linear combination of the spatial derivatives of u, and in this case the system will be called *linearly completely integrable*. It should also be remarked that the requirement that iii) hold for all complex ε can certainly be relaxed, although there seems little practical point in doing so, and that the domain Ω will customarily include the origin or at least have the origin on its boundary (it might, as in the example of the KdV equation below, be a sector of a circle center the origin).

Example 2.2. *The Korteweg–de Vries equation.* This is the original example of the use of inverse scattering techniques [21], [22]. The equation is

$$u_t+6uu_x+u_{xxx}=0, \tag{2.5}$$

and has a Lax representation with operators

$$L=-D^2-u, \qquad B=-\{4D^3+3(Du+uD)\}, \tag{2.6}$$

where $D=d/dx$. The Gel'fand–Levitan equation takes the form

$$K(x,y;t)+F(x+y;t)+\int_x^\infty K(x,z;t)F(z+y;t)\,dz=0, \tag{2.7}$$

and we recover the solution of the KdV equation via

$$u(x,t)=2\frac{d}{dx}K(x,x;t). \tag{2.8}$$

The kernel F is given by

$$F(x,t)=\sum_{j=1}^{n} c_j\exp\left(8k_j^3t-k_jx\right)+\frac{1}{2\pi}\int_{-\infty}^{\infty} R(k)\exp(2kx+8ik^3t)\,dk, \tag{2.9}$$

where $\lambda_j=-k_j^2$ are the eigenvalues, c_j the corresponding norming constants and $R(k)$ the reflection coefficient associated with the potential $u(x,0)=f(x)$. This solution is valid provided

$$\int_{-\infty}^{\infty}(1+x^2)|f(x)|\,dx<\infty \tag{2.10}$$

(cf. [7], [21]).

The uniqueness of the solution of (2.7) in the KdV case is a standard result, and the only item remaining to be checked is Definition 2.1 in condition iii). So far as analyticity is concerned, the only part of F that could fail to be analytic is that corresponding to the continuous spectrum of L:

$$F_c(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} R(k)\exp[8ik^3t+ikx]\,dk. \tag{2.11}$$

If we take any reasonable space of initial data for $\mathfrak{B}$, for example that given by (2.10), then $R(k)$ can be extended analytically into the upper half of the k-plane, and $|R(k)/k^2|$ is bounded as $|k|\to\infty$. (The function R is closely related to the spectral density function m of Titchmarsh, and the analyticity and estimates can be obtained by suitably translating the results in [33, Chap. V].) If therefore we write

$$F_c(x,t)=\frac{1}{2\pi}\left\{\int_{-\infty}^{0}+\int_{0}^{\infty}\right\}R(k)\exp[8ik^3t+ikx]\,dk=F_2+F_1,$$

say, and consider F_1, then, if t is real and positive, we can deform the integral from $(0,\infty)$ to $(0,\alpha\infty)$, for any α with $0<\arg\alpha<\frac{1}{3}\pi$. We can now increase $\arg t$, but the range for α becomes $0<\arg\alpha<\frac{1}{3}(\pi-\arg t)$. Nonetheless this does allow us to define $F_1(x,t)$ as an analytic function of t for $0<\arg t<\pi$. (It is also an analytic function of x since for large k the term k^3t dominates kx.) If we decrease $\arg t$, the range for α becomes $-\frac{1}{3}\arg t<\arg\alpha<\frac{1}{3}\pi$, which allows us to define $F_1(x,t)$ as an analytic function of t for $-\pi<\arg t<0$, and so in fact in the whole complex plane cut along the negative axis. Similar remarks apply to F_2.

Further, by using the deformed contours and integrating by parts (integrating e^{ikx} and differentiating the remainder), we see that $F\in\mathfrak{B}$, the Banach space defined by (2.10), and that the operator T is compact in $\mathfrak{B}$, although $\mathfrak{B}$ is certainly not the only possible choice for $\mathfrak{B}^*$.

Example 2.3. A case in which the combination $Q(u)$ appearing in the definition 2.1 of complete integrability is nontrivial is provided by the sine-Gordon equation, which is

$$u_{xt}=\sin u. \tag{2.12}$$

The scattering problem which can be used to solve the sine-Gordon equation was first described by Zakharov and Shabat [39] and was developed in full detail by Ablowitz, Kaup, Newell and Segur [1]; it takes the form

$$\begin{aligned} v_x&=-i\zeta v-\tfrac{1}{2}u_x(x,t)w,\\ w_x&=i\zeta w+\tfrac{1}{2}u_x(x,t)v, \end{aligned} \tag{2.13}$$

in which the x-derivative u_x of the solution of the sine-Gordon equation appears as a potential.

The analogue of the Gel'fand–Levitan equation for (2.13) again takes the form (2.7), but in this case K and F are now 2×2 matrices of functions. The matrix F is constructed from the appropriate scattering data for (2.13); the precise details of this construction can be found in [1]. Since u_x appears as the potential in (2.13), the analogue of (2.8), used to recover the solution of the sine-Gordon equation, takes the form

$$u_x(x,t)=-2K_{12}(x,x;t),$$

where K_{12} denotes the upper right-hand entry of the matrix K. Thus for the sine-Gordon equation, $Q(u)=u_x$, a fact that will be of significance when we analyze the travelling wave solutions in §3.

We now investigate the properties of the solutions of a general integral equation of Gel'fand–Levitan type. Our main tool is the following theorem of Steinberg [32], generalizing a theorem of Dolph, McLeod and Thoe [9], for the case of Hilbert–Schmidt operators.

THEOREM 2.4. *Let $\mathcal{B}$ be a Banach space, and let $T(z)$ be an analytic family of compact operators defined for $z\in\Omega\subset\mathbf{C}$. Then either $I-T(z)$ is nowhere invertible for $z\in\Omega$ or $(I-T(z))^{-1}$ is meromorphic for $z\in\Omega$.*

Let us write the Gel'fand–Levitan equation (2.3) in the symbolic form

$$(I+T(x,t))K(x,y;t)+F(x,y;t)=0, \tag{2.14}$$

where $T(x,t)$ denotes the family of integral operators

$$T(x,t)f(y)=\int_x^\infty f(z)H(z,y;t)\,dz. \tag{2.15}$$

It will always be assumed that $T(x,t)$ is a compact operator for each fixed (x,t). For instance, this is guaranteed if

$$\int_x^\infty\int_x^\infty |H(y,z;t)|^2\,dy\,dz<\infty;$$

indeed, in this case T is Hilbert–Schmidt.

To apply Steinberg's theorem, we treat the time t as the complex parameter. (Note that it would not do any good to look at x as this parameter since the domain of integration for $T(x,t)$ depends on x, and so the operators could not possibly be analytic for a large enough class of functions.) Now, for all x,y, if the kernel $H(x,y;t)$ depends analytically on t for $t\in\Omega$, then the operators $T(x,t)$ depend analytically on t. If furthermore $F(x,y;t)$ is analytic in t, then Steinberg's theorem implies that

$$K(x,y;t)=-(I+T(x,t))^{-1}F(x,y;t)$$

is, for each fixed (x,y), a meromorphic function of t. (It is one of the assumptions of complete integrability that the inverse exists for at least one t.) Therefore

$$Q[u(x,t)]=P[K(x,x;t)]$$

is also a meromorphic function of t for each fixed x.

THEOREM 2.5. *If a system of partial differential equations is Q-completely integrable in the Banach space $\mathcal{B}$, and if the initial data $u(x,0)\in\mathcal{B}$, then the function $Q[u(x,t)]$ is meromorphic in t for $t\in\Omega$ and each fixed x.*

A slight generalization of this theorem will prove to be of use in the sequel. Suppose that the time axis is "skewed", by making the change of variables

$$(\tilde{x},\tilde{t})=(x+\varepsilon t,t)$$

for some real ε. If $u=f(x,t)$ is the solution to the "unskewed" equation, then $\tilde{u}=\tilde{f}(\tilde{x},\tilde{t})=f(\tilde{x}-\varepsilon\tilde{t},\tilde{t})$ is the solution in terms of the new coordinates. If we let

$$\tilde{K}(\tilde{x},\tilde{y};\tilde{t})=K(\tilde{x}-\varepsilon\tilde{t},\tilde{y}-\varepsilon\tilde{t};\tilde{t}),$$

then $\tilde{K}$ is a solution of a Gel'fand–Levitan equation of the form

$$K(\tilde{x},\tilde{y};\tilde{t})+F(\tilde{x}-\varepsilon\tilde{t},\tilde{y}-\varepsilon\tilde{t};\tilde{t})+\int_{\tilde{x}}^\infty \tilde{K}(\tilde{x},\tilde{z};\tilde{t})H(\tilde{z}-\varepsilon\tilde{t},\tilde{y}-\varepsilon\tilde{t};\tilde{t})\,d\tilde{z}=0.$$

Therefore the "skewed" equation is also completely integrable, which gives the following theorem.

THEOREM 2.6. *If a system of partial differential equations is Q-completely integrable in the Banach space $\mathcal{B}$, and if the initial data are in $\mathcal{B}$, then the function $Q[u(x,t)]$ is meromorphic in (x,t) for $x\in\mathbf{C}$, $t\in\Omega$.*

Consider now the particular solutions of a given completely integrable system which are invariant under the action of a one-parameter symmetry group of the system. In many examples, the group is either a group of translations, leading to travelling wave solutions, or a group of scale transformations, leading to the self-similar solutions of dimensional analysis. The theory for more general symmetry groups is no more difficult than for these particular well-known examples, but in order to preserve the continuity of the exposition, we relegate a brief overview of the theory of group invariant solutions of partial differential equations to an appendix. More comprehensive treatments may be found in [6], [30] and [27].

The main result required here, which is standard for the two main examples, is that, roughly speaking, all solutions invariant under a p-parameter group G of symmetries of a given system $\Delta=0$ of partial differential equations can be found by integrating a system $\Delta/G=0$ of differential equations involving p fewer independent variables. For example, if $\Delta=0$ is a single equation for the function $u(x,t)$, $x,t\in\mathbf{R}$, the solutions invariant under the translation group G_c: $(x,t,u)\to(x+c\varepsilon,t+\varepsilon,u)$, $\varepsilon\in\mathbf{R}$, where c, the wave speed, is fixed, are just the travelling wave solutions

$$u=w(\xi), \qquad \xi=x-ct,$$

obtained as solutions of an ordinary differential equation found by substituting the above expression into the given equation. Similarly, a scaling group G_s: $(x,t,u)\to(\lambda^\alpha x,\lambda^\beta t,\lambda^\gamma u)$, $0<\lambda\in\mathbf{R}$ has self-similar solutions of the form

$$u=t^{\gamma/\beta}w(\xi), \qquad \xi=x/t^{\alpha/\beta},$$

again obtained as solutions of an ordinary differential equation.

We now state the precise hypotheses required to prove our version of the general conjecture on completely integrable systems and Painlevé type equations. For a definition of terms the reader should consult the Appendix.

We restrict our attention to a Q-completely integrable system, $\Delta=0$, of partial differential equations in two independent variables (x,t). Let G be a one-parameter local projectable symmetry group of the given system, such that the transformations in G, when extended to complex values of the variables (x,t,u), are analytic. Let G_0 denote the projected group action on (x,t)-space. Assume further that the action of G_0 on some subdomain $D_0\subset\mathbf{C}\times\Omega$, Ω as in Definition 2.1, is regular in the sense of Palais [31], so that all the G-invariant solutions of $\Delta=0$ defined over D_0 are found by integrating a system of ordinary differential equations, $\Delta/G=0$, defined over the image M_0 of D_0 in the quotient manifold M.

THEOREM 2.7. *Suppose $\Delta=0$ is a Q-completely integrable system of partial differential equations in the Banach space $\mathcal{B}$ with an analytic, regular, projectable, one-parameter symmetry group G. If $u=f(x,t)$ is a G-invariant solution of $\Delta=0$ with initial data lying in $\mathcal{B}$, then the combination corresponding to Q of the solution of the reduced system of ordinary differential equations is meromorphic in M_0, the image of D_0 in M.*

Proof. Since G_0 is analytic, the orbits of G_0 in the (x,t)-plane must be analytic curves. If the solution of the reduced equation had a singularity other than a pole on M_0, the corresponding G-invariant solution would have a similar singularity along the orbit corresponding to the singular point. This, however, would contradict Theorem 2.6. □

Thus Theorem 2.7, in a certain restricted sense, states that the reduced equation for the G-invariant solutions must be of Painlevé type. However, since the initial data for the G-invariant solutions must lie in $\mathcal{B}$, it is not for every solution of the reduced

equation that Q is required to have only poles for singularities. In effect we can consider only those solutions which either decay sufficiently rapidly at $\pm\infty$ along the real axis, or are periodic along the real axis. This restriction seems inescapable given the particular method of proof. It would be extremely interesting to remove these restrictions and prove the conjecture of the introduction in full generality.

3. Applications.

3.1. The generalized KdV equations. Consider the equation

$$u_t+u^p u_x+u_{xxx}=0, \tag{3.1}$$

where p is a nonnegative integer. This equation has scale-invariant solutions, but as the resulting third order ordinary differential equation is rather complex to analyze in full, we therefore apply our results to a simpler class of self-similar solutions, namely the travelling wave solutions. Here the symmetry group is

$$G_c\colon (x,t,u)\to(x+c\varepsilon,t+\varepsilon,u), \qquad \varepsilon\in\mathbf{R},$$

where c denotes the velocity of the wave. The invariants of G_c are $\xi=x-ct, u$, and the reduced equation for G_c-invariant solutions takes the form

$$u'''+u^p u'-cu'=0,$$

primes denoting derivatives with respect to ξ. This can be integrated once:

$$u''=\frac{-1}{p+1}u^{p+1}+cu+\frac{1}{2}d.$$

Multiplying by u', a further integration yields

$$(u')^2=\frac{-2}{(p+1)(p+2)}u^{p+2}+cu^2+du+e, \tag{3.2}$$

for some constants d,e. Thus the general travelling wave solution will be expressed in terms of the hyperelliptic function corresponding to the square root of the $(p+2)$nd order polynomial on the right of (3.2). The following two results characterize the singularities of the solutions of (3.2).

THEOREM 3.1 (*Painlevé's theorem*). *Consider the ordinary differential equation*

$$G(u',u,\xi)=0,$$

where G is a polynomial in u' and u, and analytic in ξ. Then the movable singularities of the solutions are poles and/or algebraic branch-points.

THEOREM 3.2. *Consider the equation*

$$(u')^2=R(u), \tag{3.3}$$

where R is a rational function of u. Then the solutions of (3.3) *are all meromorphic in* $\mathbf{C}$ *if and only if R is a polynomial of degree not exceeding* 4.

The proofs may be found in Ince [18] and Hille [17, p. 683]. Note that if u has an algebraic branch point, so also does any linear combination of u and its derivatives. Therefore, for (3.1) to be linearly completely integrable, (3.2) must satisfy Theorem 3.2. Thus $p=0, 1$, or 2, and in these cases the solutions are given by elliptic or trigonometric functions. Note that $p=0$ corresponds to the linear case, $p=1$ to the KdV equation, and $p=2$ to the modified KdV equation, all of which are known to be integrable by inverse scattering.

To complete the demonstration that the generalized KdV equations are not linearly completely integrable for $p\neq 0,1,2$, we must place the complete integrability in a

suitable Banach space $\mathfrak{B}$, and to do so we check the asymptotic behavior of the travelling wave solutions at $\pm\infty$. If we require that $u, u_x \to 0$ as $|x| \to \infty$, then $d=e=0$ in (3.2). Moreover the polynomial on the right of (3.2) now has a double zero at $u=0$ and a simple zero at $u_0=[\frac{1}{2}(p+1)(p+2)c]^{1/p}$. Standard techniques (cf. [36]) allow us to conclude the existence of travelling wave solutions with positive velocities decaying exponentially for $|x| \to \infty$, and reaching an extreme value of u_0. Thus for p odd, the travelling waves are humps with u_0 the peak value, while for p even, both humps and troughs occur. The important point, however, is the exponential decay of these waves for $|x| \to \infty$, and the fact that for $p \neq 0, 1, 2$, they have complex nonpolar singularities. If therefore we take for the Banach space $\mathfrak{B}$ a space of functions vanishing exponentially, we have shown that the generalized KdV equations are not linearly completely integrable in $\mathfrak{B}$ for $p \neq 0, 1, 2$, and this completes the demonstration that these equations can be solved by inverse scattering only when $p=0, 1$ or 2. This result is in accordance with numerical evidence [10] that only in these special cases do the equations have soliton solutions.

3.2. Nonlinear Klein–Gordon equations. Consider the nonlinear Klein–Gordon equation in characteristic coordinates

$$u_{xt}=f'(u), \tag{3.4}$$

where f is an analytic function of u, real for real u, and prime denotes derivative. The cases we will be most interested in are when f is a polynomial or a finite sum of exponential functions. We will determine necessary conditions on f for (3.4) to be linearly completely integrable by analysis of the singularities of the travelling wave solutions. If c is the velocity, $\xi=x-ct$, then the reduced equation for the G_c-invariant solutions of (3.4) is

$$-cu''=f'(u). \tag{3.5}$$

Multiplying (3.5) by u' and integrating yields

$$-\frac{c}{2}(u')^2=f(u)+k \tag{3.6}$$

for some constant k. For simplicity we shall assume that k can be chosen so that u_1 (real) is a simple or double zero of $f(u)+k$ and there is a second consecutive simple or double zero for some real u_2. This assumption ensures that the initial data $u(x,0)$ can be chosen to lie in a suitable Banach space $\mathfrak{B}$:

i) if u_1 and u_2 are simple zeros, so that a solution of (3.6) oscillates between u_1 and u_2, we take $\mathfrak{B}$ to be a space of periodic functions;

ii) if u_1 is a double and u_2 a simple zero, so that a solution of (3.6) decays exponentially to u_1 as $|\xi| \to \infty$, we take $\mathfrak{B}$ to be a space of functions exponentially converging:

iii) if u_1 and u_2 are double zeros, so that a solution of (3.6) tends exponentially to u_1 as $\xi \to \infty$ and to u_2 as $\xi \to -\infty$ (or vice versa), we can again take $\mathfrak{B}$ to be a space of functions exponentially converging, but to different limits.

The following theorem (stated in the context of (3.4) although it applies generally) is an immediate consequence of considering a linear combination of u and its derivatives. It tells us what singularities are possible for solutions of linearly completely integrable equations.

THEOREM 3.3. *Suppose for some constant k that the analytic function $f(u)+k$ has two consecutive simple and/or double zeros on the real axis. Then, if the nonlinear Klein–Gordon equation* (3.4) *is linearly completely integrable in the relevant Banach space*

indicated above, it must be the case that any solution of (3.6) (*with c having the opposite sign to* $f(u)+k$ *between the zeros*) *has as singularities only poles or logarithmic branch-points.*

A logarithmic branch-point is by definition a singularity such that some linear combination of derivatives has a pole. It arises in practice if the scattering operator L depends only on $u_x, u_{xx}, \cdots$, so that $Q[u]$ in turn depends only on derivatives; to demonstrate that this situation can indeed arise, consider the sine-Gordon equation

$$u_{xt}=\sin u.$$

It was indicated in Example 2.3 that this is completely integrable, and to examine it in the context of Theorem 3.3 we take

$$f(u)=-\cos u, \qquad k=0.$$

The solution of (3.6) is then

$$\sqrt{2}\sin(\tfrac{1}{2}u)=\operatorname{sn}\{c^{-1/2}(\xi+\delta)\},$$

where sn is the Jacobi elliptic function with modulus $k=1/\sqrt{2}$. This is well defined for $c>0$. Now sn has simple poles on a certain rectangular lattice in $\mathbf{C}$, and so u has logarithmic singularities at these lattice points. The reason for the appearance of these nonpolar singularities is the fact that u_x rather than u appears in the scattering operator L. We note that u_x on the other hand does have only poles for singularities.

THEOREM 3.4. *Suppose that* $f(u)$ *is a rational function, real for real u and such that, for some* k, $f(u)+k$ *has two consecutive simple and/or double zeros on the real axis. If the Klein–Gordon equation* $u_{xt}=f'(u)$ *is linearly completely integrable, then f is a polynomial of degree not exceeding* 4.

The proof is immediate from Theorems 3.1 and 3.2.

To discuss the case where f is a polynomial of degree ≤ 4, one can try other similarity solutions of (3.4), or else quite different tests. For example, it can be shown [8] that when f is of degree >2, so that f' is nonlinear, (3.4) has only finitely many polynomial conservation laws, while a theorem of Gel'fand and Dikii [12], [13] states that if a system of partial differential equations has a Lax representation, then there are an infinite number of polynomial conservation laws. Next we consider the case where f is a finite sum of exponential functions

$$f(u)=\sum_{j=0}^{m} c_j e^{\alpha_j u}, \qquad c_j, \alpha_j \in \mathbf{C}.$$

For simplicity, we restrict our attention to the case where $\alpha_j=n_j\alpha$ for some $\alpha\in\mathbf{C}$ and some rational numbers n_j. By dividing α by the common denominator of the n_j, we may assume the n_j are integers. Now let $v=\exp(\alpha u)$, so that $v'=\alpha v u'$. Thus v satisfies

$$-\frac{c}{2\alpha^2}(v')^2=\sum c_j v^{n_j+2}. \tag{3.7}$$

Note that Theorem 3.2 cannot be applied here since v may have singularities not shared by u. However, since $u'=v'/\alpha v$, it is necessary to find conditions on (3.7) such that the function v'/v, for solutions v, has no movable algebraic branch-points. This requires a more detailed investigation of the proof of Theorem 3.2. It suffices for our purposes to note the following:

LEMMA 3.5. *Consider the ordinary differential equation*

$$(v')^2=v^{-n}P(v),$$

where P is a polynomial with $P(0)\neq 0$ and n is a positive integer. Then for any $\xi_0\in\mathbf{C}$ there is a solution v with algebraic branch-point at ξ_0. This solution has a Puiseux expansion

$$v(\xi)=\sum_{j=1}^{\infty} a_j(\xi-\xi_0)^{jr}$$

with $a_1\neq 0$, and the rational number r is given by

i) $r=(m+1)^{-1}$ *if* $n=2m$,

ii) $r=2(2m+3)^{-1}$ *if* $n=2m+1$.

The proof of this result can be inferred from Hille, [17, pp. 681–682].

LEMMA 3.6. *Suppose v has an algebraic branch-point at ξ_0. Then v'/v has no branch-point at ξ_0 if and only if $v(\xi)=(\xi-\xi_0)^r f(\xi)$ for r rational and f meromorphic at ξ_0.*

Proof. Assume without loss of generality that $\xi_0=0$. Let v have the Puiseux expansion

$$v(\xi)=\xi^{mr}\sum_{j=0}^{\infty} a_j\xi^{jr},$$

where m is an integer and $a_0\neq 0$. Let a_k be the first nonzero coefficient for which kr is not an integer, if such exists. Now

$$\frac{1}{v}=\xi^{-mr}\sum_{j=0}^{\infty} b_j\xi^{jr},$$

where $b_0=a_0^{-1}$ and the first nonzero coefficient b_j with jr not an integer is $b_k=-a_k a_0^{-2}$. Furthermore

$$v'=\xi^{mr-1}\sum_{j=0}^{\infty}(m+j)ra_j\xi^{jr}.$$

Therefore

$$\frac{v'}{v}=\xi^{-1}\sum_{j=0}^{\infty} c_j\xi^{jr}$$

and the coefficient of ξ^{kr} is

$$c_k=-mra_k a_0^{-1}+(m+k)ra_k a_0^{-1},$$

which vanishes only when $a_k=0$. This proves the lemma. □

PROPOSITION 3.7. *Consider the ordinary differential equation*

$$(v')^2=\sum_{j=-n}^{N} b_j v^j. \tag{3.8}$$

Given $\xi_0\in\mathbf{C}$, there exists a solution v of (3.8) *such that v'/v has an algebraic branch-point at ξ_0, unless* (3.8) *is of the special form*

$$(v')^2=\sum_{j=-2}^{2} c_j v^{jk+2} \tag{3.9}$$

for some integer k.

Proof. Let $\xi_0=0$ and assume $b_N\neq 0$, $b_{-n}\neq 0$. By Lemma 3.6 all solutions must be of the form $v(\xi)=\xi^r f(\xi)$ with r rational and f meromorphic at 0 if we are to avoid an algebraic branch-point for v'/v. Thus

$$(v')^2=\xi^{2r}(r\xi^{-1}f+f')^2,$$

and

$$v^j=\xi^{jr}f^j,$$

so that, equating the fractional powers of ξ, we see that $b_j=0$ unless $jr=2r+\iota$ for some integer ι. If $n>0$, it follows from Lemma 3.5 that $b_j=0$ unless

i) $j\equiv 2 \bmod(m+1)$ for $n=2m$, or

ii) $2j\equiv 4 \bmod(2m+3)$ for $n=2m+1$.

In particular, the only negative values of j which satisfy these congruences are $1-\frac{1}{2}n$ and $-n$, the first value occurring only when n is even.

Next set $w=1/v$. Then (3.8) becomes

$$(w')^2=\sum_{j=-n}^{N} b_j w^{4-j}.$$

Since $w'/w=-v'/v$, w must satisfy the same conditions as v. Therefore, if $N>4$, $b_j=0$ unless

i) $j\equiv 2 \bmod(M-1)$ if $N=2M$, or

ii) $2j\equiv 4 \bmod(2M-2)$ if $N=2M+1$.

The only positive values of j satisfying these are N, $\frac{1}{2}N+1$ and 2, the second only if N is even. Comparison of the two sets of congruences then shows that (3.8) must be of the required form. □

THEOREM 3.8. *Suppose $f(u)$ is a linear combination of exponential functions $e^{\alpha_j u}$ with $\alpha_j=n_j\alpha$, n_j rational, α complex. Suppose further that $f(u)$ is real for u real, and that, for some real k, $f(u)+k$ has two consecutive simple and/or double zeros on the real axis. If the Klein–Gordon equation $u_{xt}=f'(u)$ is linearly completely integrable, then f must be of the special form*

$$f(u)=\sum_{j=-2}^{2} c_j e^{j\beta u}, \tag{3.10}$$

where β is a rational multiple of α.

It is interesting that the form (3.10) for f includes the double sine-Gordon equation

$$u_{xt}=a\sin\alpha u+b\sin\left(\frac{1}{2}\alpha u\right),$$

for which numerical studies of Dodd and Bullough [8] indicate the existence of soliton solutions. Mikhailov [24], [25] and Fordy and Gibbons [15], have shown that a special case of (3.10) when $f(u)=e^{2u}+e^{-u}$ does have a Lax representation, but it is not known whether the result extends to a general function $f(u)$ of the form (3.10).

3.3. Model wave equations of Whitham and Benjamin. The integro-differential equation

$$u_t+uu_x+\mathcal{H}[u_x]=0, \tag{3.11}$$

where $\mathcal{H}$ is the integral operator

$$\mathcal{H}[f](x)=\int_{-\infty}^{\infty} H(x-y)f(y)\,dy,$$

was proposed by Whitham [35], [36] as an alternative to the KdV equation for long waves in shallow water which could also model breaking and peaking. Here $\mathcal{K}$ is taken to be the Fourier transform of the desired phase velocity $c(k)$, where k is the wave number. Of particular interest is the case

$$c(k)=\frac{1}{\nu^2+k^2}, \qquad \nu>0,$$

so that

$$H(x)=\frac{1}{2\nu}e^{-\nu|x|}.$$

Note that $\mathcal{K}$ is the Green's function of the operator $D^2-\nu^2=\mathcal{D}$ so that (3.11) is equivalent to the differential equation

$$\mathcal{D}[u_t+uu_x]+u_x=0. \tag{3.12}$$

It can be shown [10] that (3.12) possesses travelling wave solutions u, with $|u|\to 0$ as $|x|\to\infty$, and amplitudes between 0 and some maximum height. Computer studies indicate that these waves may be solitons, i.e., they may interact cleanly. One possibly undesirable feature of (3.11) is the extremely fast propagation of short-wave components, and for this reason Benjamin, Bona and Mahony [5] proposed the alternative model

$$u_t+uu_x-\mathcal{K}[u_t]=0. \tag{3.13}$$

Again, in the special case, (3.13) can be rewritten as

$$\mathcal{D}[u_t+uu_x]-u_t=0. \tag{3.14}$$

In general, we will let $\mathcal{D}$ be any constant coefficient linear differential operator

$$\mathcal{D}=\sum_{i=0}^{n} c_i D^i, \qquad c_n\neq 0.$$

We show here that the model equations (3.12), (3.14) cannot be integrable by inverse scattering methods. As usual, consider the travelling wave solutions of these equations. If c denotes the velocity, then the reduced equation, after integration, is

$$\mathcal{D}\left[\frac{1}{2}(u-c)^2\right]+a(u+d)=0. \tag{3.15}$$

Here d is a constant of integration, $a=1$ in the Whitham model, $a=c$ in the Benjamin model, and D now denotes $d/d\xi$, $\xi=x-ct$. Since nth order equations of Painlevé type have not been classified, we resort to Painlevé's original "α-method" to analyze the singularities of the solutions of (3.15). The basic result is found in Ince [18, p. 319].

LEMMA 3.9. *Suppose* $\Delta(u,\xi,\alpha)=0$ *is an analytically parametrized family of ordinary differential equations for* α *in some domain* Ω *containing* 0 *as an interior point. If the general solution* $u(\xi,\alpha)$ *is uniform in* ξ *for* $\alpha\in\Omega\sim\{0\}$, *then it will be uniform for* $\alpha=0$.

In our case, let $\xi=\xi_0+\alpha\zeta$. Then if we consider u as a function of ζ, (3.15) becomes

$$\left(c_n D^n+\alpha c_{n-1}D^{n-1}+\cdots+\alpha^n c_0\right)\left[\frac{1}{2}(u-c)^2\right]+\alpha^n a(u+d)=0,$$

where D now denotes $d/d\zeta$. For $\alpha=0$, this reduces to

$$D^n\left(\frac{1}{2}(u-c)^2\right)=0,$$

the solution of which is

$$u=c+\sqrt{P_n(\zeta)}$$

for an arbitrary polynomial P_n of degree $\leq n-1$. This, for appropriate P_n, has an algebraic branch-point at $\zeta=0$, so that, by the lemma, solutions of (3.15) must also have nonlogarithmic branch-points. (This involves a slight extension of the lemma above, but it is easy to infer its truth from the proof given by Ince.) If these solutions also satisfy decay or periodicity properties, Theorem 2.7 (together with Theorem 3.1) shows that model equations (3.12), (3.14) cannot be linearly completely integrable. In particular, Whitham's equation with $\mathfrak{D}=D^2-\nu^2$ is not integrable by inverse scattering.

3.4. The BBM equation. The equation

$$u_t+uu_x-u_{xxt}=0, \tag{3.16}$$

known as the BBM equation, was proposed by Benjamin, Bona and Mahony [5] as an alternative model to the KdV equation for the description of long waves in shallow water. In [29] it was shown to possess only three independent conservation laws, and therefore by the results of Gel'fand and Dikii cannot be completely integrable. Our consideration of this example runs into difficulties because the self-similar solutions do not satisfy any decay or periodicity properties, and the functions Q we can allow are limited, but we will indicate the method here.

First we note that (3.16) admits the symmetry group

$$G\colon (x,t,u)\to(x,\lambda^{-1}t,\lambda u), \qquad 0<\lambda\in\mathbf{R},$$

of scale transformations. Invariants of G are provided by x and $w=tu$, for $t>0$, and the reduced equation for G=invariant solutions is then

$$w''+ww'-w=0, \tag{3.17}$$

the primes denoting derivatives with respect to x. It can be readily checked, by the procedure in Ince [18], that (3.17) is not of Painlevé type. Indeed, it is of Ince's type i(b) [18, p. 330]. Applying the α-method as Ince does, one can readily check that branch-points appear, although possibly only logarithmic, and this, granted the existence of a suitable Banach space $\mathfrak{B}$, would show that the BBM equation is not Q-completely integrable for Q, say, the identity.

However, a closer investigation of the behavior of the real solutions of (3.17) is required. Since x does not appear, it can be integrated to yield

$$(1-w')e^{w'}=ce^{-w^2/2}. \tag{3.18}$$

In principle, this equation can again be integrated by solving for w' in terms of w. To investigate the solutions qualitatively, note that $w'=0$ if and only if $w^2=2\log c$, $c\geq 1$. The only double root is when $c=1$, and only in this case do solutions decay at $+\infty$ or $-\infty$. However, it is readily seen that a solution decaying at one endpoint cannot decay at the other, nor are periodic solutions possible. Thus we are unable to apply our results to this case.

3.5. Lax pairs of composite order. Gel'fand and Dikii [12], [13] succeeded in classifying all Lax pairs of differential operators of the following special type. Let

$$L_n=D^n+u_{n-2}D^{n-2}+\cdots+u_1D+u_0$$

be a scalar differential operator of order n with $u=(u_0,\cdots,u_{n-2})$ independent C^∞ functions, and $D=d/dx$. They showed that for each integer m not a multiple of n, there is a differential operator

$$P_m=D^m+p_{m,m-2}D^{m-2}+\cdots+p_{m,1}D+p_{m,0}$$

of order m, the $p_{m,i}$ being polynomials in the u_j and their derivatives, such that the Lax representation

$$\frac{\partial L_n}{\partial t}=[P_m,L_n]$$

is a nontrivial system of evolution equations

$$u_t=K_m(u). \tag{3.19}$$

Moreover, the P_m are unique if we require the coefficients $p_{m,j}$ to have no constant term.

Consider the stationary solutions of the system (3.19), i.e., those in which u is independent of t. These satisfy the system $K_m(u)=0$, or equivalently, the "stationary Lax representation"

$$[P_m,L_n]=0. \tag{3.20}$$

THEOREM 3.10. *If the orders n,m of the operators L_n,P_m in the Lax representation of* (3.19) *are not relatively prime integers, then stationary solutions of* (3.19) *with arbitrary singularities in the complex plane exist.*

Proof. Let $k>1$ be the greatest common divisor of m and n. Consider the operator

$$M_k=D^k+v_{k-2}D^{k-2}+\cdots+v_1D+v_0,$$

whose coefficients $v_j(x)$ are sufficiently differentiable for $x\in\mathbf{R}$ but are otherwise arbitrary functions. Then

$$L_{n,0}=(M_k)^{n/k},\qquad P_{m,0}=(M_k)^{m/k}$$

obviously satisfy the stationary Lax representation (3.20) and, moreover, using the formalism of Gel'fand and Dikii, it is easy to prove that $P_{m,0}$ is derivable from $L_{n,0}$ via the same formulae as gave P_m from L_n. Therefore each such M_k gives a stationary solution of the evolutionary system (3.19). □

Now suppose that L_n is any such operator, where n is a composite number. If there exists a Gel'fand–Levitan type of integral equation for solving the inverse problem for the operator L_n, then Theorem 2.7 would imply the meromorphic character of the group-invariant solutions of the evolutionary system (3.19), using similar arguments to those used in the integration of the Korteweg–de Vries equation. This, however, is in contradiction to Theorem 3.10 for the case of time-invariant solutions. (The relevant symmetry group is just translation in t.) This indicates that such a differential operator of composite order does not have an inverse-scattering formalism in the sense that the Schrödinger operator does—either no such Gel'fand–Levitan equation exists, or the assumptions regarding analyticity are not justified. Indeed, we know of no such Gel'fand–Levitan equation for any operator of composite order, e.g. for order $n=4$.

Appendix. Group-invariant solutions of differential equations. The general theory was developed by Lie and, more recently, Ovsjannikov. For details, the best references are [6], [27], [30]. Here we briefly review the relevant concepts.

Let

$$\Delta(x,u)=0,\qquad x\in\mathbf{R}^m,\quad u\in\mathbf{R}^n, \tag{A1}$$

be a system of partial differential equations in m independent and n dependent variables. A *symmetry group* is a local Lie group of transformations acting on the space $\mathbf{R}^n \times \mathbf{R}^m$ which takes solutions of the system to other solutions. (The group acts on solutions by transforming their graphs. In the case of *projectable groups*, meaning that all transformations are of the form $(\tilde{x},\tilde{u})=(\alpha(x),\beta(x,u))$, a solution $u=f(x)$ will be transformed into the solution $\tilde{u}=\tilde{f}(\tilde{x})=\beta(\alpha^{-1}(\tilde{x}),f(\alpha^{-1}(\tilde{x})))$, provided α is invertible.)

The most helpful property of continuous symmetry groups is that for a given system they can *all* be found by systematic computations of an elementary character. The key step, which was Lie's fundamental discovery, is to look for the infinitesimal generators of the group, which are vector fields of the general form

$$v=\sum_{i=1}^{m} \xi^i(x,u)\frac{\partial}{\partial x_i}+\sum_{j=1}^{n} \varphi_j(x,u)\frac{\partial}{\partial u_j},$$

the group transformations themselves being recovered from the auxiliary ordinary differential equations governing the integration of the above vector field. This leads to the following infinitesimal criterion for a symmetry group of a given system [28].

THEOREM. *Let G be a connected local Lie group. Then G is a symmetry group of the system of partial differential equations $\Delta=0$ if and only if*

$$\text{(A2)} \qquad \operatorname{pr} v(\Delta)=0 \quad \textit{whenever } \Delta=0$$

for every infinitesimal generator v of G.

Here $\operatorname{pr} v$ refers to the "prolongation" of the vector field v, obtained as the infinitesimal generator of the action of the group G on the spaces of partial derivatives of u with respect to x induced by the action of G on functions $u=f(x)$. The point is that the condition (A2) leads to a large number of elementary partial differential equations for the coefficients ξ^i, φ_j of v, the general solution of which is the most general infinitesimal generator of a one-parameter symmetry group of the given system of differential equations. Examples of this computation can be found in the above-mentioned references.

Now, given a symmetry group G, a G-invariant (or self-similar) solution of (A1) is a solution which is unchanged by the transformations in G. The fundamental property of G-invariant solutions is that, roughly speaking, they may all be found via the integration of a system of partial differential equations in fewer independent variables. To make this precise, we must assume that G acts "regularly" in the sense of Palais [31] on an open subset $U\subset \mathbf{R}^m\times\mathbf{R}^n$. This requires, in U,

i) that all the orbits of G have the same dimension,

ii) that, for any point (x,u), there exist arbitrarily small neighborhoods N such that the intersection of any orbit O of G with N is a connected subset of O.

(The prototypical group actions excluded by the second requirement are the irrational flows on the torus.)

Under these two assumptions, it is well known that the quotient space $M=U/G$, whose points correspond to the orbits of G, can be naturally endowed with the structure of a smooth (although not always Hausdorff) manifold. Moreover, the G-invariant solutions of (A1) are all obtained by integrating a reduced system $\Delta/G=0$ of partial differential equations on M, which necessarily has fewer independent variables. Precise statements and proofs of these results may be found in [27].

For our purposes, the construction of the reduced system for the G-invariant solutions proceeds as follows: Local coordinate systems on the quotient manifold M are

provided by a "complete set of functionally independent invariants of G", cf. [30]. If G is projectable, these are functions of the form

$$\xi^1(x),\cdots,\xi^{m-l}(x),w^1(x,u),\cdots,w^n(x,u),$$

which are unchanged under the action of G. The functional independence means that the Jacobian matrix

$$\begin{pmatrix} \partial\xi/\partial x & 0 \\ \partial w/\partial x & \partial w/\partial u \end{pmatrix}$$

is everywhere nonsingular. The reduced system $\Delta/G=0$ will then be found in terms of the new independent variables ξ^i and the new dependent variables w^j.

REFERENCES

[1] M. J. ABLOWITZ, D. J. KAUP, A. C. NEWELL AND H. SEGUR, *The inverse scattering transform–Fourier analysis for nonlinear problems*, Stud. Appl. Math., 53 (1974), pp. 249–315.

[2] M. J. ABLOWITZ, A. RAMANI AND H. SEGUR, *Nonlinear evolution equations and ordinary differential equations of Painlevé type*, Lett. Nuovo Cimento, 23 (1978), pp. 333–338.

[3] _____, *A connection between nonlinear evolution equations and ordinary differential equations of P-type*. I, J. Math. Phys., 21 (1980), pp. 715–721.

[4] M. J. ABLOWITZ AND H. SEGUR, *Exact linearization of a Painlevé transcendent*, Phys. Rev. Lett., 38 (1977), pp. 1103–1106.

[5] T. B. BENJAMIN, J. L. BONA AND J. J. MAHONY, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[6] G. W. BLUMAN AND J. D. COLE, *Similarity Methods for Differential Equations*, Lecture Notes in Appl. Math. Sci., 13, Springer-Verlag, New York, 1974.

[7] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.

[8] R. K. DODD AND R. K. BULLOUGH, *Bäcklund transformations for the sine-Gordon equations*, Proc. Roy. Soc. London Ser. A, 351 (1976), pp. 499–523.

[9] C. L. DOLPH, J. B. MCLEOD AND D. THOE, *The analytic continuation of the resolvent kernel and scattering operator associated with the Schrödinger operator*, J. Math. Anal. Appl., 16 (1966), pp. 311–332.

[10] B. FORNBERG AND G. B. WHITHAM, *A numerical and theoretical study of certain nonlinear wave phenomena*, Philos. Trans. Roy. Soc. London Ser. A, 289 (1978), pp. 373–404.

[11] C. S. GARDNER, *Korteweg–de Vries equation and generalizations*. IV. *The Korteweg–de Vries equation as a Hamiltonian system*, J. Math. Phys., 12 (1971), pp. 1548–1551.

[12] I. M. GEL'FAND AND L. A. DIKII, *Fractional powers of operators and Hamiltonian systems*, Functional Anal. Appl., 10 (1976), pp. 259–273.

[13] _____, *Resolvents and Hamiltonian systems*, Functional Anal. Appl., 11 (1977), pp. 93–105.

[14] I. M. GEL'FAND AND B. M. LEVITAN, *On the determination of a differential equation for its spectral function*, Amer. Math. Soc. Transl., Ser. 2, 1 (1955), pp. 253–304.

[15] J. GIBBONS AND A. P. FORDY, *A class of integrable nonlinear Klein–Gordon equations in many dependent variables*, preprint, Dublin Institute for Advanced Studies.

[16] S. P. HASTINGS AND J. B. MCLEOD, *A boundary value problem associated with the second Painlevé transcendent and the Korteweg–de Vries equation*, Arch. Rational Mech. Anal., 73 (1980), pp. 31–51.

[17] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, London, 1968.

[18] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1944.

[19] D. J. KAUP, *The three-wave interaction—a nondispersive phenomenon*, Stud. Appl. Math., 55 (1976), pp. 9–44.

[20] _____, *On the inverse scattering problem for cubic eigenvalue problems of the class* $\Psi_{xxx}+q\Psi_x+r\Psi=\lambda\Psi$, Stud. Appl. Math., 62 (1980), pp. 189–216.

[21] G. L. LAMB, *Elements of Soliton Theory*, Wiley, New York, 1980.

[22] P. D. LAX, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[23] H. P. MCKEAN AND E. TRUBOWITZ, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29 (1976), pp. 143–226.

[24] A. V. MIKHAILOV, *Integrability of a two-dimensional generalization of the Toda chain*, Soviet Phys. JETP Lett. 30 (1979), pp. 414–418.

[25] _____, *The reduction problem and the inverse scattering method*, Physica, 3D (1981), pp. 73–117.

[26] R. M. MIURA, C. S. GARDNER AND M. D. KRUSKAL, *Korteweg–de Vries equation and generalizations. II. Existence of conservation laws and constants of motion*, J. Math. Phys., 9 (1968), pp. 1204–1209.

[27] P. J. OLVER, *Symmetry groups and group invariant solutions of differential equations*, J. Differential Geom., 14 (1979), pp. 497–542.

[28] _____, *How to find the symmetry group of a differential equation*, appendix in D. H. Sattinger, Group Theoretic Methods in Bifurcation Theory, Lecture Notes in Math. 762, Springer-Verlag, New York, 1979.

[29] _____, *Euler operators and conservation laws of the BBM equation*, Math. Proc. Cambridge Philos. Soc., 85 (1979), pp. 143–160.

[30] L. V. OVSJANNIKOV, *Group Properties of Differential Equations*, Novosibirsk, 1962 (translated by G. W. Bluman, unpublished).

[31] R. S. PALAIS, *A Global Formulation of the Lie Theory of Transformation Groups*, Memoirs 22, American Mathematical Society, Providence, RI, 1957.

[32] S. STEINBERG, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1969), pp. 372–379.

[33] E. C. TITCHMARSH, *Eigenfunction Expansions Associated with Second-order Differential Equations*, 2nd ed., Oxford Univ. Press, Oxford, 1962.

[34] E. TRUBOWITZ, *The inverse problem for periodic potentials*, Comm. Pure Appl. Math., 30 (1977), pp. 321–337.

[35] G. B. WHITHAM, *Variational methods and applications to water waves*, Proc. Roy. Soc. London Ser. A, 299 (1967), pp. 6–25.

[36] _____, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

[37] V. E. ZAKHAROV AND S. V. MANAKOV, *Resonant interaction of wave packets in nonlinear media*, Soviet Phys. JETP Lett., 18 (1973), pp. 243–245.

[38] _____, *On the complete integrability of the nonlinear Schrödinger equation*, Zh. Matem. i Teor. Fiz., 19 (1974), pp. 322–343.

[39] V. E. ZAKHAROV AND A. B. SHABAT, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Phys. JETP, 34 (1972), pp. 62–69.

[40] _____, *A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem*, Functional Anal. Appl., 8 (1974), pp. 226–235.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0009 $01.25/0

A NUMERICAL TREATMENT FOR PARABOLIC EQUATIONS WITH A SMALL PARAMETER*

GEORGE C. HSIAO† AND KIRK E. JORDAN‡

Abstract. A modified Crank–Nicolson–Galerkin procedure is developed for treating initial-boundary value problems for parabolic equations with a small parameter, multiplying the time-derivative term. Error analyis as well as numerical experiments are included. It is shown that with rather moderate step sizes, the numerical results are in excellent agreement with the theoretical ones both inside and outside the initial layer.

1. Introduction. Let Ω denote a bounded domain with a smooth boundary $\partial\Omega$ in R^n and let $T>0$ be any fixed constant. Consider the initial-boundary value problem (P_ε) consisting of the parabolic equation

$$\varepsilon\frac{\partial u}{\partial t}+L[u]=f(x,t), \qquad (x,t)\in\Omega\times(0,T], \tag{E}$$

together with the initial condition

$$u(x,0)=\mathring{u}(x), \qquad x\in\bar{\Omega}, \tag{I}$$

and the homogeneous boundary condition

$$u(x,t)=0, \qquad (x,t)\in\partial\Omega\times[0,T], \tag{B}$$

where $\varepsilon>0$ is a small parameter, f and $\mathring{u}$ are given smooth functions satisfying certain regularity conditions to be specified. The operator L is a strongly elliptic second-order partial differential operator with C^∞ coefficients of the form

$$L[u]:=-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right)+c(x)u, \tag{1.1}$$

satisfying the ellipticity condition

$$\sum_{i,j=1}^{n}a_{ij}(x)\xi_i\xi_j\geq\lambda_0\sum_{i=1}^{n}\xi_i^2, \tag{1.2}$$

with a constant $\lambda_0>0$ for any real vector $(\xi_1,\xi_2,\cdots,\xi_n)\in R^n$ and any point $x\in\bar{\Omega}$. In addition, we assume that $a_{ij}=a_{ji}$ and $c\geq 0$ in $\bar{\Omega}$.

Problem (P_ε) represents an important class of singular perturbation problems, which are normally investigated by the method of matched asymptotic expansions or the method of composite expansions [4], [12]. In terms of the terminology in the singular perturbation theory, there is an initial layer in the neighborhood of $t=0$. Outside this layer, the solution of (P_ε) is mainly dominated by that of the reduced problem, a boundary value problem in the present case,

$$\begin{aligned} L[U]&=f(x,t), \qquad x\in\Omega,\\ U&=0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{P_0}$$

* Received by the editors September 25, 1981, and in revised form May 10, 1982.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711.

‡ Exxon Research and Engineering Company, Computing Technology and Services Division, Linden, New Jersey 07036. This research was conducted in part while the second author was Assistant Professor of Mathematics, State University of New York at Oswego, Oswego, New York 13126.

for each fixed t, $0 \le t \le T$. Clearly the solution $U(x,t)$ of (P_0) generally is not expected to satisfy the initial condition (I), and hence one is led to consider the *initial-layer problem*:

$$(\tilde{P}_0) \qquad \begin{aligned} &\frac{\partial V}{\partial \tilde{t}} + L[V] = 0, \qquad x \in \Omega, \quad \tilde{t} > 0, \\ &V(x,0) = \mathring{u}(x) - U(x,0), \qquad x \in \bar{\Omega}, \\ &V = 0, \qquad x \in \partial\Omega, \quad \tilde{t} \ge 0, \end{aligned}$$

where $\tilde{t} := t/\varepsilon$ is the *stretched variable*. As a typical feature of the singular perturbation problems, the solution $V = V(x,\tilde{t})$ is significant only within the initial layer and decays exponentially as $\tilde{t} \to \infty$. Then the exact solution $u = u(x,t;\varepsilon)$ of (P_ε), in appropriate function spaces, admits an asymptotic expansion of the form (see [6]):

$$u(x,t;\varepsilon) = U(x,t) + V(x,\tilde{t}) + Z(x,t;\varepsilon) \tag{1.3}$$

where the remainder term $Z = O(\varepsilon)$ as $\varepsilon \to 0^+$ *uniformly* for all t, $0 \le t \le T$. Here U and V are only the leading terms in the asymptotic expansion for u, and higher order terms are defined similarly by the reduced and initial-layer problems such as (P_0) and $(\tilde{P}_0)$.

From the singular perturbation results, one may approximate the exact solution by the solutions of the reduced and initial-layer problems for small ε, if the solutions U and V are available. However in the present case, both (P_0) and $(\tilde{P}_0)$ cannot be solved explicitly and hence one has to rely on some approximate schemes. On the other hand, one may be tempted to solve the singular perturbation problem (P_ε) directly by the standard numerical scheme such as the Crank–Nicolson–Galerkin scheme, since (E) is a linear parabolic equation for each $\varepsilon > 0$, no matter how small. As will be seen, because of the presence of ε, the usual Crank–Nicolson–Galerkin schemes can not be directly applied and one will not obtain meaningful numerical results without reducing mesh size in the initial layer. This of course requires considerable computational effort. In fact, often because of the limitation of the computer system, the discrete problems involved may become ill-posed numerically when mesh sizes get to be too small.

In this paper, we present two numerical procedures for treating singular perturbation problems such as (P_ε). Neither one of them needs very fine mesh size. In essence, our procedures use singular perturbation theory to construct the leading terms in the formal asymptotic expansions (1.3). Motivated by the asymptotic behavior of the singular perturbation problems, we solved (P_ε) numerically via the reduced and initial-layer problems by the Galerkin method with finite elements as trial functions for the space variables. This leads to an initial-value problem for a system of ordinary differential equations with constant coefficients and hence explicit solutions can be constructed. We refer to this approximation simply as the *Galerkin approximation*. For the purpose of implementation on a computer, the system of ordinart differential equations in the Galerkin approximation is further discretized and we arrive at a fully discrete method for (P_ε) which we refer to as the *Initial-Layer-Crank–Nicolson–Galerkin* (ILCNG) *approximation*. We comment that although (E) is considered only for the finite time interval $0 \le t \le T$, in terms of stretched variable $\tilde{t}$, it is really an infinite time interval, $0 \le \tilde{t} \le T/\varepsilon$ for ε sufficiently small. As will be seen, the simple idea introduced in [7] as how to determine a reasonable approximate finite domain with respect to the stretched variable can be adopted here very naturally.

We organize the paper as follows. In §2, we formulate and describe the approximate schemes in detail. Sections 3 and 4 contain the error estimates for the approximations. Finally, in §5, a simple example is included to show the applicability of the

ILCNG approximation. Clearly with very little computational effort the numerical results obtained from our scheme are definitely far better than the ones from applying directly the standard Crank–Nicolson–Galerkin scheme to the problem, especially within the initial layer.

2. Numerical procedures. Since our numerical procedures are based on the Galerkin's method, it is most appropriate to consider the weak formulations of the problems (P_ε), (P_0) and $(\tilde{P}_0)$. First we need some notation. As usual, we denote by $H^m(\Omega)$ the real Sobolev space of order m (an integer) on Ω equipped with the norm $\|\cdot\|_m$ and $\mathring{H}^m(\Omega)$ the subspace of $H^m(\Omega)$ obtained by completing $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_m$; $C_0^\infty(\Omega)$ denotes the space of infinitely differentiable functions with compact support in Ω.

By a weak solution of (P_ε), we mean a function $u=u(x,t)$ such that for each fixed $t\in[0,T]$, $u(t):=u(\cdot,t)\in\mathring{H}^1(\Omega)$ and satisfies the integral identities:

$$\begin{aligned}&\left(\varepsilon\frac{\partial u(t)}{\partial t},w\right)+a(u(t),w)=(f(t),w)\quad\text{for all } w\in\mathring{H}^1(\Omega),\quad t\in(0,T],\\&(u(0)-\mathring{u},w)=0\quad\text{for all } w\in\mathring{H}^1(\Omega).\end{aligned}\tag{2.1}$$

Here $(\cdot,\cdot)$ denotes the L_2-inner product and $a(\cdot,\cdot)$ is the bilinear form associated with L,

$$a(u,w):=\int_\Omega\left\{\sum_{i,j=1}^n a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial w}{\partial x_j}+c(x)uw\right\}dx.\tag{2.2}$$

Similarly, for each fixed $t\in[0,T]$, $U(t):=U(\cdot,t)\in\mathring{H}^1(\Omega)$ is the weak solution of the reduced problem (P_0) satisfying

$$a(U(t),w)=(f(t),w)\quad\text{for all } w\in\mathring{H}^1(\Omega),\quad t\in[0,T],\tag{2.3}$$

and $V(\tilde{t})=V(\cdot,\tilde{t})\in\mathring{H}^1(\Omega)$, the weak solution of the initial-layer problem $(\tilde{P}_0)$ such that

$$\begin{aligned}&\left(\frac{\partial V(\tilde{t})}{\partial\tilde{t}},w\right)+a(V(\tilde{t}),w)=0\quad\text{for all } w\in\mathring{H}^1(\Omega),\quad \tilde{t}>0,\\&(V(0)+U(0)-\mathring{u},w)=0\quad\text{for all } w\in\mathring{H}^1(\Omega),\end{aligned}\tag{2.4}$$

where $\tilde{t}=t/\varepsilon$ is the stretched variable.

For simplicity we assume that f is sufficiently smooth so that u, U and V are continuously differentiable with respect to t. More precisely, if $C^k(I;H)$ denotes H-valued functions which are k times continuously differentiable on the interval I, we require that u and U belong to $C^1((0,T];\mathring{H}^1(\Omega))\cap C^0([0,T];\mathring{H}^1(\Omega))$ and that $V=V(x,\tilde{t})$ belongs to $C^1((0,T/\varepsilon];\mathring{H}^1(\Omega))\cap C^0([0,T/\varepsilon];H^1(\Omega))$. The existence and uniqueness of the weak solutions for (P_ε), $(\tilde{P}_0)$ and (P_0) follow from the standard results for linear parabolic and elliptic equations (see, e.g. [10]). In particular, it is well known that under the assumptions on the coefficients, the bilinear form $a(\cdot,\cdot)$ in (2.2) is continuous on $H^1(\Omega)\times H^1(\Omega)$ and strongly coercive on $\mathring{H}^1(\Omega)$ for every $t\in[0,\infty)$. That is, there are positive constants λ and μ such that

$$|a(u,v)|\le\mu\|u\|_1\|v\|_1\quad\text{for all } u,v\in H^1(\Omega)\tag{2.5}$$

and

$$a(u,u)\ge\lambda\|u\|_1^2\quad\text{for all } u\in\mathring{H}^1(\Omega).\tag{2.6}$$

These properties are equally crucial to the error bounds for our numerical approximations.

To describe *the Galerkin approximation*, let us denote by S^h a one-parameter family of $N(h)$-dimensional subspaces of $\mathring{H}^1(\Omega)$ with a certain interpolation property to be specified later. We assume that $\{\phi_k(x)\}_{k=1}^N$ forms a basis of S^h and propose the following approximation:

$$u_*^h := U^h + V^h \tag{2.7}$$

where U^h and V^h are the continuous-in-time Galerkin approximations for U and V respectively in (2.3) and (2.4). In terms of the basis $\{\phi_k(x)\}_{k=1}^N$, the approximations U^h and V^h have the explicit representations:

$$U^h = \sum_{k=1}^N \alpha_k(t)\phi_k(x) \quad \text{and} \quad V^h = \sum_{k=1}^N \beta_k(\tilde{t})\phi_k(x), \tag{2.8}$$

Here α_k's are solutions of the algebraic system:

$$\sum_{k=1}^N a(\phi_k,\phi_l)\alpha_k(t) = (f(t),\phi_l), \qquad l=1,2,\cdots,N \tag{2.9}$$

which is equivalent to the Galerkin's equation corresponding to (2.3). The β_k's are the solutions of the initial-value problem for the system of ordinary equations from the Galerkin equation corresponding to (2.4):

$$\begin{aligned} &\sum_{k=1}^N \{(\phi_k,\phi_l)\dot{\beta}_k(\tilde{t}) + a(\phi_k,\phi_l)\beta_k(\tilde{t})\} = 0, \qquad \tilde{t}>0, \\ &\sum_{k=1}^N (\phi_k,\phi_l)\beta_k(0) = \left(\mathring{u} - \sum_{k=1}^N \alpha_k(0)\phi_k, \phi_l\right) \end{aligned} \tag{2.10}$$

for $l=1,2,\cdots,N$. Clearly both (2.9) and (2.10) are uniquely solvable, since the matrices appearing in (2.9) and (2.10) are all positive definite. Indeed, by denoting $A=a(\phi_k,\phi_l)$ the stiffness matrix and $B=(\phi_k,\phi_l)$ the Gramm matrix it is easily seen that the approximation u_*^h defined by (2.7) admits the explicit representation:

$$u_*^h = \sum_{k=1}^N \gamma_k^0(t,\tilde{t})\phi_k(x)$$

with $\gamma_k^0(t,\tilde{t})$ defined by

$$\gamma_k^0(t,\tilde{t}) = \left(A^{-1}\mathbf{f}(t) + e^{-\tilde{t}\Lambda}B^{-1}\mathbf{u}_0\right)\cdot \hat{e}_k \tag{2.11}$$

where $\hat{e}_k = k$th unit vector in $\mathbb{R}^N$, $\mathbf{f}(t) = ((f(t),\phi_1),\cdots,(f(t),\phi_N))^T$,

$$\mathbf{u}_0 = ((\mathring{u},\phi_1),\cdots,(\mathring{u},\phi_N))^T - B^{-1}(\alpha_1(0),\cdots,\alpha_N(0))^T,$$
$$\Lambda = B^{-1}A.$$

We comment that if we solve (2.1) directly by the Galerkin method and approximate u by

$$u^h = \sum_{k=1}^N \gamma_k(t)\phi_k(x),$$

then $\gamma=(\gamma_1(t),\cdots,\gamma_N(t))^T$ will be the solution of the singular perturbation problem for the ordinary differential equation:

$$\begin{aligned} &\varepsilon B\gamma'(t)+A\gamma(t)=\mathbf{f}(t), \qquad 0<t\le T,\\ &\gamma(0)=B^{-1}((\mathring{u},\phi_1),\cdots,(\mathring{u},\phi_N))^T. \end{aligned} \tag{2.12}$$

In terms of the terminology in the singular perturbation theory, it is easy to see that the vector $\gamma^0=(\gamma_1^0,\cdots,\gamma_N^0)^T$ defined by (2.11) is really just the zeroth order term in the composite expansion for γ in (2.12). Hence our *Galerkin approximation* is a combination of *asymptotic* and *numerical* approximations. The coefficients γ_k^0 defined by (2.11) are computable in principle. The exponential term in the coefficient is clearly the analytic solution of the initial-value problem (2.10). In practice, it is best handled by difference approximations or Padé approximations. This leads us to the *Initial-Layer-Crank–Nicolson–Galerkin* (ILCNG) *approximation*, a fully discrete procedure which is readily implemented on the computer.

Before we describe the ILCNG approximation, we note that since the reduced problem (2.3) contains no ε, it can be treated by any standard scheme for the boundary value problem, e.g. the Galerkin method. On the other hand, although one may handle the initial-layer problem (2.4) by any explicit scheme, it requires a long time to reach $t=T$, since at $t=T$, $\tilde{t}=T/\varepsilon$ will be rather large for ε small. To circumvent this difficulty, following [7] we solve (2.4) only for $0\le\tilde{t}\le\tilde{m}$, where $\tilde{m}$ is chosen such that

$$e^{-\lambda\tilde{m}}\le\varepsilon \quad \text{or} \quad \tilde{m}\ge\frac{\ln\varepsilon}{\lambda}, \tag{2.13}$$

where λ is the coercivity constant in (2.6). Naturally this choice of $\tilde{m}$ is based on the asymptotic behavior of γ in (2.12). A simple computaton shows that the solution $\tilde{\boldsymbol{\beta}}(\tilde{t})=(\beta_1(\tilde{t}),\cdots,\beta_N(\tilde{t}))^T$ of (2.10) and the coefficient $\gamma^0(t,\tilde{t})=(\gamma_1^0(t,\tilde{t}),\cdots,\gamma_N^0(t,\tilde{t}))^T$ of (2.11) satisfy the estimates:

$$\begin{aligned} &|\boldsymbol{\beta}(\tilde{t})|\le\sqrt{\kappa(B)}\,|\boldsymbol{\beta}(0)|e^{-\lambda\tilde{t}} \quad \text{for all } \tilde{t}\ge0,\\ &|\gamma(t)-\gamma^0(t,\tilde{t})|=O(\varepsilon) \quad \text{as } \varepsilon\to0^+ \end{aligned} \tag{2.14}$$

uniformly for all $0\le t\le T$, where $\kappa(B)$ is the condition number of B and $|\cdot|$ stands for the usual Euclidean norm. Or from (2.1), (2.3) and (2.4), similar estimates hold also for u, U and V; that is,

$$\begin{aligned} &\|V(\tilde{t})\|_0\le e^{-\lambda\tilde{t}}\|V(0)\|_0 \quad \text{for all } \tilde{t}\ge0,\\ &\|u(t)-U(t)-V(\tilde{t})\|_0=O(\varepsilon) \quad \text{as } \varepsilon\to0^+ \end{aligned} \tag{2.14'}$$

uniformly for all $t\in[0,T]$. Hence in view of (2.14) and (2.14'), if one is interested in an approximation only up to $O(\varepsilon)$, the choice of $\tilde{m}$ in (2.13) is not unreasonable. In this way we have modified the problem (2.4) and consider (2.4) only for $0\le\tilde{t}\le\tilde{m}$.

Our ILCNG approximation can now be described simply as a procedure consisting of the Galerkin method for (2.3) and the usual Crank–Nicolson–Galerkin method for (2.4) with $0\le\tilde{t}\le\tilde{m}$. To be more specific, we approximate u of (2.1) by

$$\tilde{u}_*^h(x,t):=\begin{cases} U^h(x,t), & t>\tilde{m}\varepsilon,\\ U^h(x,t)+V_j^h, & t=\varepsilon\tilde{t}_j\le\tilde{m}\varepsilon. \end{cases} \tag{2.15}$$

Here U^h is the Galerkin solution of (2.3) which is the same as in the case of the continuous-in-time and is defined by (2.8) and (2.9), while V_j^h's are the Crank–Nicolson–Galerkin solutions of (2.4) at $\tilde{t}_j=j\Delta\tilde{t}$, $j=0,1,\cdots,J=\tilde{m}/\Delta\tilde{t}$ with $\Delta\tilde{t}$ denoting the mesh size in the stretched time $\tilde{t}$-direction. In terms of the basis $\{\phi_k(x)\}_{k=1}^N$ of S^h, V_j^h admits the representation:

$$V_j^h=\sum_{k=1}^{N}\beta_{kj}\phi_k(x),\qquad j=0,1,\cdots,J=\tilde{m}/\Delta\tilde{t}, \tag{2.16}$$

where $\boldsymbol{\beta}_j=(\beta_{1j},\beta_{2j},\cdots,\beta_{Nj})^T$ are defined explicitly by the recurrence relations [13]:

$$\begin{aligned}&\boldsymbol{\beta}_0:=B^{-1}\mathbf{u}_0,\\ &\boldsymbol{\beta}_{j+1}:=\left(I+\tfrac{1}{2}\Delta\tilde{t}B^{-1}A\right)^{-1}\left(I-\tfrac{1}{2}\Delta\tilde{t}B^{-1}A\right)\boldsymbol{\beta}_j\end{aligned} \tag{2.17}$$

for $j=0,1,\cdots,J-1$, where A,B and $\mathbf{u}_0$ have the same meanings as those given in (2.11). It is clear that all the $\boldsymbol{\beta}_j$'s as well as the coefficients $\alpha(t)$'s of U^h in (2.15) are now ready to be computed by routine numerical algorithms since in these calculations the small parameter ε does not appear explicitly.

In order to show the applicability of the ILCNG approximation, numerical experiments are performed on model problems in §5. As will be seen, with very little computational effort the numerical results obtained from our scheme are definitely far better than the ones from applying directly the Crank–Nicolson–Galerkin scheme to the problem, especially within the initial layer. Both Galerkin approximation and ILCNG approximation procedures are easily extended to the general case where higher order terms in the asymptotic expansions are included. However, in order to achieve optimal rate of convergence, the accuracy of the numerical approximation should be varied according to the order of ε. This will become clear from the error estimates of our approximations in the next two sections.

3. The Galerkin approximation. In this section we derive the error estimates for the Galerkin approximation, a continuous-time Galerkin procedure. Since the accuracy of the procedure depends on the properties of the approximating subspaces S^h, following [1] we assume that for fixed integer $m\geq 2$, the finite-dimensional subspaces S^h of $\mathring{H}^1(\Omega)$ possess the approximation property such that for any $u\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$, $1\leq s\leq m$,

$$\inf_{\chi\in S^h}\{\|u-\chi\|_0+h\|u-\chi\|_1\}\leq ch^s\|u\|_s \tag{3.1}$$

holds, where c is a constant independent of h and u. As is well known, the space of piecewise linear polynomials satisfies (3.1) for $s=2$. In general we have the following error estimates.

THEOREM 3.1. *Let $\{S^h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces of $\mathring{H}^1(\Omega)$ satisfying the approximation property* (3.1). *Let u_*^h be the Galerkin approximation defined by* (2.7). *For smooth data, $\mathring{u}\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$ and $\frac{\partial f}{\partial t}\in H^0(\Omega)$, if $u\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$ is the exact solution of* (2.1) *such that $\frac{\partial u}{\partial t}\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$, then the following estimate*

$$\begin{aligned}\left\|u(t)-u_*^h(t)\right\|_0\leq{}& ch^s\left\{e^{-\lambda t/\varepsilon}\|\mathring{u}\|_s+\|u(t)\|_s\right\}\\ &+\frac{\varepsilon}{\lambda}(1-e^{-\lambda t/\varepsilon})\left\{Ch^s\sup_{0<\tau\leq t}\left\|\frac{\partial u}{\partial t}(\tau)\right\|_s+\frac{1}{\lambda}\sup_{0<\tau\leq t}\left\|\frac{\partial f}{\partial t}(\tau)\right\|_0\right\}\end{aligned} \tag{3.2}$$

holds for $0\leq t\leq T$, where C is a constant independent of ε,h and u.

Remark. The estimate (3.2) implies that for given ε, $h_{\text{opt}}=\varepsilon^{1/s}$ is an optimal choice of h in the sense that the error is $O(\varepsilon)$ uniformly for $0\le t\le T$.

We begin the proof of (3.2) by splitting the error estimate of $u-u_*^h$ into

$$u-u_*^h=(u-\tilde{u})+(\tilde{u}-u^h)+\left(u^h-u_*^h\right) \tag{3.3}$$

and consider each term separately. Here u^h denotes the Galerkin solution of (2.1) (see (2.12)), and $\tilde{u}=Pu$ is the energy projection of u onto S^h and is defined for each fixed $t\ge 0$ by

$$a(\tilde{u}(t),v)=a(u(t),v) \quad \text{for all } v\in S^h. \tag{3.4}$$

The size of $u-\tilde{u}$ is known from the approximate property (3.1). In fact it is easily shown that if for some constant $k\ge 0$, $u\in C^k((0,T];H^s(\Omega)\cap \mathring{H}^1(\Omega))$, then

$$\left\|\left(\frac{\partial}{\partial t}\right)^k(u-\tilde{u})(t)\right\|_0\le Ch^s\left\|\left(\frac{\partial}{\partial t}\right)^k u(t)\right\|_s \tag{3.5}$$

for some constant C independent of h and u, and $1\le s\le m$. The remaining terms in (3.3) are estimated in the following lemmas. We remark that the Galerkin solution u^h is utilized here only as an intermediate artifice in order to derive the estimates.

LEMMA 3.1. *If* $e(t)=\tilde{u}(\cdot,t)-u^h(\cdot,t)$, *then*

$$\varepsilon\left(\frac{\partial e}{\partial t},e\right)+a(e,e)=\varepsilon\left(\frac{\partial\tilde{u}}{\partial t}-\frac{\partial u}{\partial t},e\right). \tag{3.6}$$

Moreover, for $0\le t\le T$, *the following estimate holds*:

$$\|e(t)\|_0\le e^{-\lambda t/\varepsilon}\|e(0)\|_0+\frac{\varepsilon}{\lambda}(1-e^{-\lambda t/\varepsilon})\sup_{0<\tau\le t}\left\|\frac{\partial\tilde{u}}{\partial t}-\frac{\partial u}{\partial t}(\tau)\right\|_0, \tag{3.7}$$

provided $u\in C([0,T];H^0(\Omega))$ *and* $\partial u/\partial t\in C((0,T];H^0(\Omega))$.

Proof. By the definition, $e(t)=\tilde{u}(\cdot,t)-u^h(\cdot,t)\in S^h$, we may put $v=e$ in (2.1) as well as in the corresponding Galerkin equation to (2.1). Equation (3.6) then follows immediately from (3.4) with $v=e$, where we have tacitly used the fact that the energy projection P commutes with the differentiation (see [4]).

From (3.6) the rate of change of the error $e(t)$ is easy to find. The coercivity of $a(\cdot,\cdot)$ implies

$$a(e,e)\ge\lambda\|e\|_1^2\ge\lambda\|e\|_0^2.$$

Also note that

$$\varepsilon\left(\frac{\partial e}{\partial t},e\right)=\frac{\varepsilon}{2}\frac{d}{dt}\|e\|_0^2=\varepsilon\|e\|_0\frac{d}{dt}\|e\|_0.$$

The right-hand side of (3.6) is bounded by $\varepsilon\|\partial\tilde{u}/\partial t-\partial u/\partial t\|_0\|e\|_0$. Cancelling the common factor $\|e\|_0$, the identity (3.6) leads to

$$\varepsilon\frac{d}{dt}\|e\|_0+\lambda\|e\|_0\le\varepsilon\left\|\frac{\partial u}{\partial t}-\frac{\partial u}{\partial t}\right\|_0$$

and hence,

$$\frac{d}{dt}\left(e^{\lambda t/\varepsilon}\|e\|_0\right)\le e^{\lambda t/\varepsilon}\left\|\frac{\partial\tilde{u}}{\partial t}-\frac{\partial u}{\partial t}\right\|_0.$$

Integrating from 0 to t, we obtain

$$\|e(t)\|_0 \le e^{-\lambda t/\varepsilon}\|e(0)\|_0 + \int_0^t e^{-\lambda(t-\tau)/\varepsilon}\left\|\left(\frac{\partial \tilde{u}}{\partial t} - \frac{\partial u}{\partial t}\right)(\tau)\right\|_0 d\tau$$

$$\le e^{-\lambda t/\varepsilon}\|e(0)\|_0 + \frac{\varepsilon}{\lambda}(1-e^{-\lambda t/\varepsilon}) \sup_{0<\tau\le t}\left\|\left(\frac{\partial \tilde{u}}{\partial t} - \frac{\partial u}{\partial t}\right)(\tau)\right\|_0$$

which is the desired result (3.7).

LEMMA 3.2. *If* $\delta(t) = u^h(\cdot,t) - u_*^h(\cdot,t)$, *then*

$$\varepsilon\left(\frac{\partial \delta}{\partial t}, \delta\right) + a(\delta,\delta) = -\varepsilon\left(\frac{\partial u^h}{\partial t}, \delta\right) \tag{3.8}$$

and

$$\|\delta(t)\|_0 \le e^{-\lambda t/\varepsilon}\|\delta(0)\|_0 + \frac{\varepsilon}{\lambda^2}(1-e^{-\lambda t/\varepsilon}) \sup_{0<\tau\le t}\left\|\frac{\partial f}{\partial t}(\tau)\right\|_0 \tag{3.9}$$

for $0 \le t \le T$.

The proof of Lemma 3.2 is almost identical to that of Lemma 3.1, if one notices that $\partial U^h/\partial t$ is dominated by $\partial f/\partial t$, or more precisely,

$$\left\|\frac{\partial U^h}{\partial t}(t)\right\|_0 \le \frac{1}{\lambda}\left\|\frac{\partial f}{\partial t}(t)\right\|_0.$$

The details will be omitted.

To complete the proof, it remains only to consider the estimates of the initial terms in the lemmas. By the definition of u^h and u_*^h, it is clear that both $u^h(\cdot,0)$ and $u_*^h(\cdot,0)$ are the best L^2-approximation of $\mathring{u}$ in S^h. Hence by the property of the uniqueness of the best L^2-approximation in a strictly convex normed linear space [2], [3], it follows that $\delta(0) = 0$. On the other hand, we have

$$\|e(0)\|_0 = \|\tilde{u}(\cdot,0) - u^h(\cdot,0)\|_0 \le \|\tilde{u}(\cdot,0) - u(\cdot,0)\|_0 + \|\mathring{u} - u^h(\cdot,0)\|_0,$$

since $\|u(\cdot,0) - \mathring{u}\|_0 = 0$ for $\mathring{u} \in \mathring{H}^1(\Omega)$. Consequently for $\mathring{u}$, $u \in H^s(\Omega) \cap \mathring{H}^1(\Omega)$, it follows from (3.5) and the approximating property of S^h that

$$\|e(0)\|_0 \le Ch^s\{\|u(\cdot,0)\|_s + \|\mathring{u}\|_s\} \tag{3.10}$$

for some constant C independent of $h, \mathring{u}$ and $u(\cdot,0)$.

This completes the proof of Theorem 3.1, if one collects the results (3.5), (3.7), (3.9) and (3.10).

Remark. Under the same hypotheses in Theorem 3.1, an error estimate in H^1-norm can also be derived. In general we have

$$\left\|u(t) - u_*^h(t)\right\|_r \le C_0 h^{s-r}\left\{\|u(t)\|_s + \|\mathring{u}\|_s + h^r\varepsilon \sup_{0<\tau\le t}\left\|\frac{\partial u}{\partial t}(\tau)\right\|_0\right\}$$

$$+ C_1\varepsilon \sup_{0<\tau\le t}\left\|\frac{\partial f}{\partial t}(\tau)\right\|_0 \quad \text{for } r = 1, 0,$$

where $C_0 = C_0(\lambda, \lambda_0, C)$ and $C_1 = C_1(\lambda, \lambda_0)$ are constants.

4. The ILCNG approximation. Here we consider error estimates for the ILCNG approximation described in §2. Again we assume that S^h possesses the approximation property (3.1) as in the continuous-in-time case. We first note that the ILCNG approximation $\tilde{u}_*^h$ defined in (2.15),

$$\tilde{u}_*^h(x,t)=\begin{cases} U^h(x,t), & t>\tilde{m}\varepsilon, \\ U^h(x,t)+V_j^h, & t=\varepsilon\tilde{t}_j\le\tilde{m}\varepsilon \end{cases}$$

is essentially the numerical approximation for the zeroth order term, $U+V$ in the asymptotic expansion (1.3). It is natural to decompose the error according to

$$u-\tilde{u}_*^h=[u-(U+V)]+\left[U+V-\tilde{u}_*^h\right].$$

As will be seen, the first term yields the asymptotic error from the singular perturbation theory while the second term contains the numerical error. Indeed, if $Z:=u-(U+V)$, it is easily seen that Z is the weak solution of the initial-boundary value problem:

$$\begin{aligned} &\varepsilon\left(\frac{\partial Z}{\partial t},w\right)+a(Z,w)=-\varepsilon\left(\frac{\partial U}{\partial t},w\right) \quad \text{for all } w\in\mathring{H}^1(\Omega), \quad t\in(0,T], \\ &(Z,w)=0 \quad \text{for all } w\in\mathring{H}^1(\Omega). \end{aligned} \tag{4.1}$$

Hence in the same manner as (3.9) one can show that

$$\|Z(t)\|_0\le\frac{\varepsilon}{\lambda^2}(1-e^{-\lambda t/\varepsilon})\sup_{0<\tau\le T}\left\|\frac{\partial f}{\partial t}(\tau)\right\|_0 \tag{4.2}$$

for all t, $0\le t\le T$.

On the other hand, the estimate for the second term $(U+V-\tilde{u}_*^h)$ depends on the location of t. For t outside the initial layer, that is, $t>\tilde{m}\varepsilon$,

$$\left\|U+V-\tilde{u}_*^h\right\|_0\le\|U-U^h\|_0+\|V\|_0\le Ch^s\|f\|_{s-2}+\varepsilon\left(\|\mathring{u}\|_0+\frac{1}{\lambda}\|f(0)\|_0\right). \tag{4.3}$$

Here the first estimate follows from the standard arguments in finite element analysis for elliptic problems, whereas the second estimate follows from the initial layer behavior of V in (2.14′) and the choice of $\tilde{m}$ in (2.13).

For t within the initial layer, that is, $t=\varepsilon\tilde{t}_j\le\tilde{m}\varepsilon$, the estimate for $(U+V-\tilde{u}_*^h)$ is more involved; we introduce $\tilde{V}(x,\tilde{t})$ for each fixed $\tilde{t}\ge0$ the energy projection of V onto S^h defined by

$$a(\tilde{V}(\tilde{t}),v)=a(V(\tilde{t}),v) \quad \text{for all } v\in S^h, \tag{4.4}$$

and consider the estimate

$$\left\|U(\tilde{t})+V(\tilde{t})-\tilde{u}_*^h(\tilde{t})\right\|_0\le\|U(\tilde{t})-U^h(\tilde{t})\|_0+\|V(\tilde{t})-\tilde{V}(\tilde{t})\|_0+\left\|\tilde{V}(\tilde{t})-V_j^h\right\|_0 \tag{4.5}$$

for $\tilde{t}=\tilde{t}_j=j\Delta\tilde{t}$, $j=0,0,1,\cdots,J=\tilde{m}/\Delta\tilde{t}$ (see (2.16)). Both $U-U^h$ and $V-\tilde{V}$ can be estimated as (4.3) and (3.5). For the additional term $\tilde{V}-V_j^h$, we have the following lemma.

LEMMA 4.1. *Let* $\xi_j:=V_j^h-\tilde{V}(\cdot,\tilde{t}_j)$ *and* $\eta_j:=V(\cdot,\tilde{t}_j)-\tilde{V}(\cdot,\tilde{t}_j)$ *for* $j=0,1,\cdots,J$. *Then for* $V\in C^3([0,\tilde{m}];H^0(\Omega))$,

$$(\delta_{\tilde{t}}\xi_j,\xi_{j+1}+\xi_j)+\tfrac{1}{2}a(\xi_{j+1}+\xi_j,\xi_{j+1}+\xi_j)=(\delta_{\tilde{t}}\eta_j-\tau_j,\xi_{j+1}+\xi_j) \tag{4.6}$$

for $j=0,1,\cdots,J-1$ $(J\geq 1)$ *with* $\delta_{\tilde{t}}\xi_j := (\xi_{j+1}-\xi_j)/\Delta\tilde{t}$ *and* $\delta_{\tilde{t}}\eta_j := (\eta_{j+1}-\eta_j)/\Delta\tilde{t}$ *where* $\tau_j := (-(\Delta\tilde{t})^2/12)(\partial^3 V/\partial\tilde{t}^3)(\cdot,\hat{t})$, $\tilde{t}_j<\hat{t}<\tilde{t}_{j+1}$, *is the local truncation error. Furthermore, the estimate*

$$\|\xi_j\|_0 \leq \|\xi_0\|_0 + \sqrt{\frac{2\tilde{m}}{\lambda}}\left\{\sup_{0\leq k\leq j-1}\|\delta_{\tilde{t}}\eta_k\|_0 + \kappa_j(\Delta\tilde{t})^2\right\} \tag{4.7}$$

holds for $j=1,\cdots,J$ *where* κ_j *is a constant independent of* ε, h *and* $\Delta\tilde{t}$.

Proof. It is easy to verify that the approximate solution V_j^h defined by (2.16) and (2.17) satisfies, for $j=1,2,\cdots,J$,

$$\left(\delta_{\tilde{t}}V_j^h, w^h\right) + \tfrac{1}{2}a\left(V_{j+1}^h + V_j^h, w^h\right) = 0 \quad \text{for all } w^h\in S^h$$

together with

$$\left(V_0^h, w^h\right) = \left(\mathring{u} - U^h(\cdot,0), w^h\right) \quad \text{for all } w^h\in S^h.$$

Hence

$$\begin{aligned}\left(\delta_{\tilde{t}}\xi_j, w^h\right) + \tfrac{1}{2}a\left(\xi_{j+1}+\xi_j, w^h\right) &= -\left(\delta_{\tilde{t}}\tilde{V}\left(\tilde{t}_j\right), w^h\right) - \tfrac{1}{2}a\left(\tilde{V}\left(\tilde{t}_{j+1}\right) + \tilde{V}\left(\tilde{t}_j\right), w^h\right)\\ &= -\left(\delta_{\tilde{t}}\tilde{V}\left(\tilde{t}_j\right), w^h\right) - \tfrac{1}{2}a\left(V\left(\tilde{t}_{j+1}\right) + V\left(\tilde{t}_j\right), w^h\right)\end{aligned} \tag{4.8}$$

in view of (4.4). On the other hand, from (2.4) we have

$$\left(\delta_{\tilde{t}}V\left(\tilde{t}_j\right), w^h\right) + \tfrac{1}{2}a\left(V\left(\tilde{t}_{j+1}\right) + V\left(\tilde{t}_j\right), w^h\right) = \left(\tau_j, w^h\right) \tag{4.9}$$

where

$$\tau_j := -\frac{1}{12}(\Delta\tilde{t})^2\frac{\partial^3 V}{\partial\tilde{t}^3}(\cdot,\hat{t}), \qquad \tilde{t}_j<\hat{t}<\tilde{t}_{j+1} \tag{4.10}$$

is the local truncation error. It follows from (4.8) and (4.9) that

$$\left(\delta_{\tilde{t}}\xi_j, w^h\right) + \tfrac{1}{2}a\left(\xi_{j+1}+\xi_j, w^h\right) = \left(\delta_{\tilde{t}}\eta_j - \tau_j, w^h\right)$$

for all $w^h\in S^h$. In particular one may put $w^h=\xi_{j+1}+\xi_j\in S^h$ and obtain the desired result (4.6).

To establish (4.7), we use the identity

$$\left(\delta_{\tilde{t}}\xi_j, \xi_{j+1}+\xi_j\right) = \frac{1}{2}\delta_{\tilde{t}}\|\xi_j\|_0^2$$

and the coercivity of $a(\cdot,\cdot)$ so that from (4.6) we have

$$\delta_{\tilde{t}}\|\xi_j\|_0^2 + \lambda\|\xi_{j+1}+\xi_j\|_0^2 \leq 2\left(\delta_{\tilde{t}}\eta_j - \tau_j, \xi_{j+1}+\xi_j\right).$$

The right-hand side is dominated by

$$2\|\delta_{\tilde{t}}\eta_j - \tau_j\|_0\|\xi_{j+1}+\xi_j\|_0 \leq \frac{1}{\lambda}\|\delta_{\tilde{t}}\eta_j - \tau_j\|_0^2 + \lambda\|\xi_{j+1}+\xi_j\|_0^2.$$

Consequently, we obtain

$$\delta_{\tilde{t}}\|\xi_j\|_0^2 \leq \frac{2}{\lambda}\left\{\|\delta_{\tilde{t}}n_j\|_0^2 + \|\tau_j\|_0^2\right\}. \tag{4.11}$$

Summing (4.11) from $j=0$ to $j=k\le J-1$ yields

$$\|\xi_{k+1}\|_0^2\le\|\xi_0\|_0^2+\frac{2\Delta\tilde{t}}{\lambda}\left\{\sum_{j=0}^{k}\left(\|\delta_{\tilde{t}}\eta_j\|_0^2+\|\tau_j\|_0^2\right)\right\}.$$

Thus,

$$\|\xi_j\|_0^2\le\|\xi_0\|_0^2+\frac{2\tilde{m}}{\lambda}\left\{\sup_{0\le k\le j-1}\|\delta_{\tilde{t}}\eta_k\|_0^2+\sup_{0\le k\le j-1}\|\tau_k\|_0^2\right\}$$

since $j\Delta\tilde{t}\le J\Delta\tilde{t}=\tilde{m}$. Hence from (4.10), the result (4.7) then follows immediately with κ_j defined by

$$\kappa_j:=\frac{1}{12}\sup_{0\le\tilde{t}\le j\Delta\tilde{t}}\left\|\frac{\partial^3 V}{\partial\tilde{t}^3}\right\|_0. \tag{4.12}$$

This completes the proof of Lemma 4.1.

Lemma 4.1 provides information concerning the size of $\tilde{V}(\cdot,\tilde{t})-V_j^h$ within the initial layer. By the definition, we see that

$$\|\xi_0\|_0\le\|U^h(0)-U(0)\|_0+\|V(0)-\tilde{V}(0)\|_0, \tag{4.13}$$

the size of which is known from approximation property (3.5) and the like. The forward difference $\|\delta_{\tilde{t}}\eta_k\|_0$ can be rewritten in terms of $\partial/\partial\tilde{t}\,\eta_k$ such that

$$\|\delta_{\tilde{t}}\eta_k\|_0\le\left\|\frac{\partial\eta_{k+1/2}}{\partial\tilde{t}}\right\|_0+\frac{(\Delta\tilde{t})^2}{24}\left\|\frac{\partial^3\eta}{\partial\tilde{t}^3}(\hat{t})\right\|_0,\qquad \tilde{t}_k<\hat{t}<\tilde{t}_{k+1}, \tag{4.14}$$

where the first term on the right is again known from approximation property similar to (3.5) and the second is of order $(\Delta\tilde{t})^2$.

Finally, collecting the results (4.2), (4.3), (4.5), (4.7), (4.13) and (4.14), we arrive at the following

THEOREM 4.1. *Let $\{S^h\}_{0<h\le 1}$ be a family of finite-dimensional subspaces of $\mathring{H}^1(\Omega)$ satisfying the approximation property* (3.1). *Let $\tilde{u}_*^h$ be* ILCNG *approximation defined by* (2.15). *Then for sufficiently smooth data f and $\mathring{u}$, if $U\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$ for $0\le t\le T$, V, $\partial V/\partial\tilde{t}\in H^s(\Omega)\cap\mathring{H}^1(\Omega)$ for $0\le\tilde{t}\le\tilde{m}$, and in addition, if $\partial U/\partial t\in H^0(\Omega)$ for $0\le t\le T$ and $V\in C^3([0,\tilde{m}];H^0(\Omega))$, then the following asymptotic rate of convergence holds*:

$$\left\|u(t)-\tilde{u}_*^h(t)\right\|_0\le\begin{cases}C_0\varepsilon+C_1h^s & \text{for } \tilde{m}\varepsilon<t\le T,\\ C_0\varepsilon+C_1h^s+C_2\sqrt{m}\left(h^s+(\Delta\tilde{t})^2\right) & \text{for } t=\varepsilon\tilde{t}_j\end{cases} \tag{4.15}$$

with $\tilde{t}_j=j\Delta\tilde{t}$, $j=0,1,\cdots,J=\tilde{m}/\Delta\tilde{t}$, where the C_i's are constants independent of ε,h and $\Delta\tilde{t}$.

Remark. In the special case when S^h is the space of piecewise linear polynomials, i.e., $s=2$, this result, (4.15), was announced in [8].

5. Numerical experiments. Here we present some numerical results for the ILCNG approximation scheme previously discussed. We consider the following model problem:

$$(\mathrm{P}_\varepsilon)\qquad\begin{aligned}&\varepsilon\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=f(x,t),\qquad 0<x<1,\quad t>0,\\ &u(x,0)=\mathring{u}(x),\qquad 0<x<1,\\ &u(0,t)=u(1,t),\qquad t>0,\end{aligned}$$

with $f(x,t)=1-3x$ and $\mathring{u}(x)=\sin\pi x-\frac{1}{2}x^2(1-x)$. The exact solution of (P_ε) can be found explicitly:

$$u(x,t)=-\tfrac{1}{2}x^2(1-x)+e^{-t\pi^2/\varepsilon}\operatorname{Sin}\pi x.$$

Here the first term on the right-hand side is the reduced solution U while the second term is the initial-layer solution V. For comparison we also solve (P_ε) by the standard Crank–Nicolson–Galerkin (CNG) scheme (see [11], [2] for more details).

For simplicity in both schemes, the ILCNG and CNG schemes, the finite dimensional subspace S^h was chosen to be the space of piecewise linear polynomials. The basis functions ϕ_k are the usual hat functions defined by

$$\phi_k(x):=\begin{cases}\dfrac{x-x_{k-1}}{h}, & x_{k-1}\le x\le x_k,\\[2ex] \dfrac{x_{k+1}-x}{h}, & x_k\le x\le x_{k+1},\\[2ex] 0 & \text{otherwise}\end{cases}$$

with uniform mesh size $h=\frac{1}{N}$. An interpolation of the initial data was used for both schemes. Thus at the space nodes, x_k, when $t=0$, the approximate solutions and the exact solutions are identical. Both schemes were implemented using single precision on the Burroughs B7700 computer at the University of Delaware.

For the ILCNG approximation scheme, $\tilde{m}$ was chosen according to (2.13). In the present case, it is easily shown that the coercivity constant λ is equal to π^2, and thus,

$$\text{(C)}\qquad \tilde{m}\ge q\ln 10/\pi^2$$

if, in particular, $\varepsilon=10^{-q}$, $q>0$. We comment that in general the coercivity constant λ depends on the spectrum of the corresponding differential operator L and the exact value may not be known so easily. However, a rough estimate of λ may suffice in the determination of the lower bound for $\tilde{m}$ and explicit knowledge of λ may not be needed.

We conducted a graphical comparison of the ILCNG and CNG schemes for $\varepsilon=10^{-2}$. At the space point $x=.5$, we plotted the solutions versus time. Our first graph contains the solution from CNG and the exact solution; see Fig. 1. Figure 2 contains the ILCNG and CNG solutions. In both the ILCNG and the CNG schemes, the space mesh is chosen to be $\frac{1}{10}$. For the CGN scheme the time mesh Δt is $\frac{1}{40}$, while for the ILCNG scheme the time mesh $\Delta\tilde{t}$ is $\frac{1}{10}$ for the initial-layer solution.

One immediately notices in Figs. 1–2 the oscillations occurring in the standard CNG scheme. This of course is due to the small parameter ε appearing in the problem. As time progresses, the effects of the initial layer, which causes the oscillations eventually damp out, but these effects certainly influence the numerical approximation far outside the initial-layer region. These disturbing oscillations are not present in the ILCNG approximation and in fact the ILCNG approximation exhibits behavior similar to the exact solution as our graphs illustrate.

For the ILCNG scheme several numerical experiments were performed. We first allowed ε to vary. Of course, we had to change $\tilde{m}$ in accordance with (5.1). The remaining parameters were held fixed. The results for $h=.1$, $\Delta\tilde{t}=.1$ are summarized in Table 1.

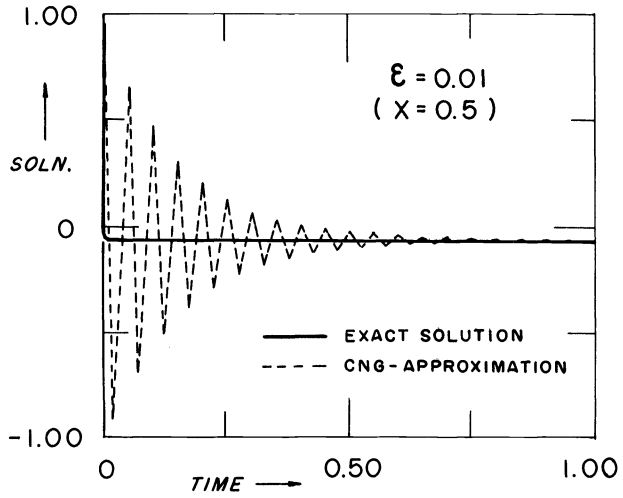


FIG. 1. CNG *scheme. Approximate solution by* CNG *scheme and exact solution.* CNG *with* $h=0.1$, $\Delta t=0.025$.

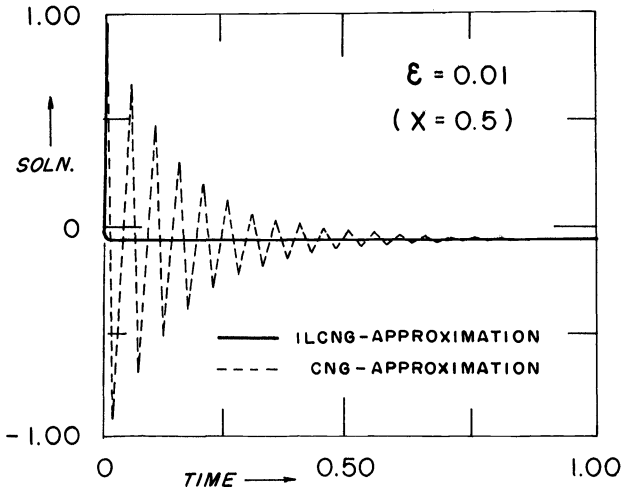


FIG. 2. ILCNG *and* CNG *overlay. Comparison of approximate solutions obtained by both schemes.* ILCNG *with* $h=0.1$, $\tilde{m}=2$, $\Delta\tilde{t}=0.1$. CNG *with* $h=0.1$, $\Delta t=0.025$.

In order to indicate the dependence on the various meshes for $\varepsilon=10^{-4}$, we varied the space and initial layer meshes, h and $\Delta\tilde{t}$ respectively. Since the maximum error occurred at $x=.5$, we tabulated the results at this space point and at the fixed time level $t=1\times10^{-5}$ in Table 2. We note that for $h=.05$ and $\Delta\tilde{t}=.01$, the approximation is almost of the same order as the small parameter ε. This indicates that by choosing $h^2=(\Delta\tilde{t})^2=\varepsilon$, we will obtain optimal order for our scheme. This is similar to results obtained in the case of ordinary differential equations reported in [7] and [9].

TABLE 1.
$h=.1, \quad \Delta\tilde{t}=.1$

t	APPROX.	EXACT	ERROR
	$\varepsilon=10^{-4}$	$\tilde{m}=2$	
.5(−4)	$-.56293\times10^{-1}$	$-.55308\times10^{-1}$	$-.98536\times10^{-3}$
.1(−3)	$-.62461\times10^{-1}$	$-.62448\times10^{-1}$	$-.13202\times10^{-4}$
.15(−3)	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	$-.13291\times10^{-6}$
.1	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	0.0
	$\varepsilon=10^{-8}$	$\tilde{m}=2$	
.5(−8)	$-.56293\times10^{-1}$	$-.55308\times10^{-1}$	$-.98536\times10^{-3}$
.1(−7)	$-.62461\times10^{-1}$	$-.62448\times10^{-1}$	$-.13202\times10^{-4}$
.15(−7)	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	$-.13291\times10^{-6}$
.1	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	0.0
	$\varepsilon=10^{-12}$	$\tilde{m}=3$	
.5(−12)	$-.56293\times10^{-1}$	$-.55308\times10^{-1}$	$-.98536\times10^{-3}$
.1(−11)	$-.62461\times10^{-1}$	$-.62448\times10^{-1}$	$-.13202\times10^{-4}$
.15(−11)	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	$-.13291\times10^{-6}$
.1	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	0.0
	$\varepsilon=10^{-16}$	$\tilde{m}=4$	
.5(−16)	$-.56293\times10^{-1}$	$-.55308\times10^{-1}$	$-.98536\times10^{-3}$
.1(−15)	$-.62461\times10^{-1}$	$-.62448\times10^{-1}$	$-.13202\times10^{-4}$
.15(−15)	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	$-.13291\times10^{-6}$
.1	$-.62500\times10^{-1}$	$-.62500\times10^{-1}$	0.0

TABLE 2
$\varepsilon=10^{-4}, x=.5, t=0.1\times10^{-4}$

$\Delta\tilde{t}$	Error when $h=.1$	Error when $h=.05$
.1	3.7195×10^{-2}	3.40339×10^{-2}
.05	1.0824×10^{-2}	8.4588×10^{-3}
.025	4.9331×10^{-3}	2.6428×10^{-3}
.0125	3.4981×10^{-3}	1.2255×10^{-3}
.01	3.3268×10^{-3}	1.0564×10^{-3}

In conclusion we emphasize that the model problem (P_ε) considered here was very simple, nevertheless it is typical. Our numerical results were rather encouraging especially in comparison with those obtained by the standard Crank–Nicolson–Galerkin method. Very little computational effort is needed for our scheme to achieve the satisfactory accuracy and obtain the precise behavior of the solution in the initial-layer region; furthermore it does not encounter the stability problems there. We hope that the idea of combining asymptotic and numerical methods employed here may shed some light on some more complicated problems.

REFERENCES

[1] I. BABUSKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 3–363.

[2] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[3] P. J. DAVIS, *Interpolation and Approximation*, Dover, New York, 1975.

[4] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam, 1979.

[5] G. FAIRWEATHER, *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, New York, 1978.

[6] F. HOPPENSTEADT, *On quasilinear parabolic equations with a small parameter*, Comm. Pure App. Math., 24 (1971), pp. 17–38.

[7] G. C. HSIAO AND K. E. JORDAN, *Solutions to the difference equations of singular perturbation problems*, in Numerical Analysis of Singular Perturbation Problems, P. W. Hemker and J. J. H. Miller, eds., Academic Press, London, 1979, pp. 433–440.

[8] _____, *A finite element method for singularly perturbed parabolic equations*, in Boundary and Interior Layers—Computation and Asymptotic Methods, J. J. H. Miller, ed., Boole Press, Dublin, 1980, pp. 317–321.

[9] K. E. JORDAN, *A numerical treatment of singularly perturbed boundary and initial-boundary value problems*, Ph.D. Dissertation, University of Delaware, Newark, DE, 1980.

[10] V. P. MIKHAILOV, *Partial Differential Equations*, MIR, Moscow, 1978.

[11] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley, New York, 1973.

[12] R. E. O'MALLEY, JR., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[13] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0010 $01.25/0

PARABOLIC CAPACITY AND SOBOLEV SPACES*

MICHEL PIERRE†

Abstract. We prove in particular here that, given an open set Ω of $\mathbb{R}^N$, the usual parabolic capacity on $]0,T[\times\Omega$ associated with the heat operator $\frac{\partial}{\partial t}-\Delta$ can be defined using only the Hilbert norm of the space $\mathcal{W}=\{v\in L^2(0,T;H_0^1(\Omega));\ \frac{\partial v}{\partial t}\in L^2(0,T;H^{-1}(\Omega))\}$ which arises in parabolic variational inequalities. The result is stated in the general setting of parabolic Dirichlet spaces.

Introduction. Let Ω be an open subset of $\mathbb{R}^N$ and $T>0$. The usual parabolic capacity on $]0,T[\times\Omega$ associated with the heat operator $\mathcal{E}=\frac{\partial}{\partial t}-\Delta$ is defined by

$$\forall\,\omega\subset]0,T[\times\Omega \text{ open},\quad c_0(\omega)=\int_{]0,T[\times\Omega} d\mathcal{E}u_\omega,$$

where u_ω is the capacitary potential of ω, that is, the solution of the (formal) variational inequality

$$\text{(I)}\qquad \begin{aligned} &u\geq 1_\omega \ \text{ a.e.},\quad u(0)=1_\omega(0),\quad u(t,\cdot)|_{\partial\Omega}=0,\\ &\frac{\partial u}{\partial t}-\Delta u\geq 0,\qquad \frac{\partial u}{\partial t}-\Delta u=0 \ \text{ on } [u>1_\omega].\end{aligned}$$

(Here 1_ω is the characteristic function of ω. Note that $\mathcal{E}u_\omega=\frac{\partial u_\omega}{\partial t}-\Delta u_\omega$ is a nonnegative measure on $]0,T[\times\Omega$.) Another definition in terms of measures can also be found in [2].

We show in this paper that this capacity can be defined using only the Hilbert norm of the space

$$\mathcal{W}=\left\{v\in L^2\big(0,T;H_0^1(\Omega)\big);\ \frac{\partial v}{\partial t}\in L^2\big(0,T;H^{-1}(\Omega)\big)\right\},$$

namely, if we set, for any open subset ω of $]0,T[\times\Omega$,

$$c(\omega)=\inf\left\{\|v\|_{\mathcal{W}}^2;\ v\geq 1_\omega \text{ a.e.}\right\},$$

where

$$\|v\|_{\mathcal{W}}^2=\|v\|_{L^2(0,T;H_0^1(\Omega))}^2+\left\|\frac{\partial v}{\partial t}\right\|_{L^2(0,T;H^{-1}(\Omega))}^2,$$

then there exist $a,b>0$ such that:

$$\text{(II)}\qquad \forall\,\omega,\quad a\cdot c_0(\omega)\leq c(\omega)\leq b\cdot c_0(\omega).$$

It is well known that this space $\mathcal{W}$ arises as the natural space of test-functions in numerous parabolic variational inequalities (V.I.) of type (I) (see Lions–Stampacchia [4], Lions–Magenes [5], Lions [3], Mignot–Puel [6],...). On the other hand, as in the elliptic case, the tools of potential theory have also proven to be most useful to solve and interpret these parabolic V.I. (see [1], [8]). The above result emphasizes the strong relationship between the two approaches.

* Received by the editors June 4, 1981. This work was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and supported in part by the National Science Foundation under grant MCS78-09525.

† Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin, 53706 Permanent address, Université Scientifique et Médicale de Grenoble, Institut de Mathématiques Pures, B.P. 116, 38402-Saint-Martin-D'Hères, France.

A direct consequence of (II) is that any element of $\mathcal{W}$ has a quasicontinuous representation. This fact (that we established in [8]) is an important tool to deduce fundamental properties about the structure of parabolic potentials (i.e., the functions $u \in L^2(0,T;H_0^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$ such that $\frac{\partial u}{\partial t} - \Delta u \geq 0$) (see [8], [10] for these results).

Another consequence is that, as in the elliptic case, "L^2-estimates" can be used to evaluate the parabolic capacity of a set. In the same spirit, we also show here the following result: if u is a parabolic potential greater than or equal to 1 on ω, then the capacity of ω can be estimated by the norm of u in $L^2(0,T;H_0^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$.

Lastly, this suggests that for the nonlinear problems associated with operators of the form

$$\frac{\partial u}{\partial t} - \operatorname{div} A(x,u,Du),$$

the natural capacity can be defined by the norm of

$$\mathcal{W}_p = \left\{ v \in L^p(0,T;W^{1,p}(\Omega)); \frac{\partial u}{\partial t} \in L^{p'}(0,T;W^{-1,p'}(\Omega)) \right\},$$

where $p \in]1,\infty[$ is suitably chosen and $\frac{1}{p} + \frac{1}{p'} = 1$.

In this paper, we state our result in the general setting of Dirichlet parabolic spaces so that it can be applied to general elliptic operators with Dirichlet, Neumann or mixed boundary conditions.

1. Parabolic Dirichlet space. Let X be a locally compact space, countable at the infinity,[1] ξ a Radon measure on X whose support is X. We denote $\mathcal{K}(X)$ (resp. $\mathcal{K}^+(X)$) the space of continuous (resp. nonnegative and continuous) real functions with compact support in X. The space $\mathcal{K}(X)$ is equipped with its usual locally convex topology.

Let V be a Hilbert space with the norm $\|\cdot\|$; we assume that V is embedded into $L^2(X)$, the space of (classes of) real square integrable functions with the norm

$$|u|_2 = \left[\int_X u^2(x)\, d\xi(x) \right]^{1/2}.$$

Then, if V' is the dual space of V, we have

$$V \hookrightarrow L^2(X) \hookrightarrow V'. \tag{1}$$

The scalar product in $L^2(X)$ as well as the duality (V', V) will be denoted by $(\cdot,\cdot)$.

We will assume:

$$\mathcal{K}(X) \cap V \text{ is dense in } V \text{ and } \mathcal{K}(X). \tag{2}$$

Example 1. (α) $X = \mathbb{R}^N$, $V = H^1(\mathbb{R}^N)$, $V' = H^{-1}(\mathbb{R}^N)$.
(β) $X = \Omega$ open set in $\mathbb{R}^N$, $V = H_0^1(\Omega)$, $V' = H^{-1}(\Omega)$.
(γ) $X = \bar{\Omega}$, $V = H^1(\Omega)$ (Ω regular bounded open set in $\mathbb{R}^N$).
(δ) $X = \{1 \text{ point}\}$, $V \approx L^2(X) \approx \mathbb{R}$.

Given $T > 0$, we denote $Q = [0,T[\times X$ equipped with the Radon measure $dt \otimes \xi$, where dt is the Lebesgue measure on $[0,T[$. $\mathcal{K}(Q)$ will denote the space of continuous numerical functions with compact support in Q, equipped with its natural topology.

[1] That is X is the union of a countable number of compact subsets.

Now, associated with V, V', we have

$$\mathcal{V}=L^2(0,T;V) \text{ and its dual } \mathcal{V}'=L^2(0,T;V'),$$

$$\mathcal{W}=\left\{v\in\mathcal{V};\ \frac{dv}{dt}\in\mathcal{V}'\right\}.$$

These spaces are Hilbert spaces with the norms:

$$\|v\|_{\mathcal{V}}^2=\int_0^T\|v(t)\|^2dt,\quad \|v\|_{\mathcal{V}'}^2=\int_0^T\|v(t)\|_{V'}^2dt,\quad \|v\|_{\mathcal{W}}^2=\|v\|_{\mathcal{V}}^2+\left\|\frac{\partial v}{\partial t}\right\|_{\mathcal{V}'}^2.$$

Let us recall that $\mathcal{W}$ is embedded into $C([0,T[;L^2(X))$ (see Lions–Magenes [5]).

As a consequence of (2), one can show that (see [8])

(3) $$\mathcal{K}(\tilde{Q})\cap\mathcal{W} \text{ is dense in } \mathcal{W} \text{ and } \mathcal{K}(\tilde{Q}),\qquad \tilde{Q}=[0,T]\times X.$$

The operators $A(t)$. For a.e. t, let $a(t,\cdot,\cdot)$ be a bilinear form on $V\times V$ satisfying:

(4) $$\forall u,v\in V\times V,\quad t\mapsto a(t,u,v) \text{ is measurable},$$

(5) $$\exists M\geq 0,\quad \forall (u,v)\in V\times V,\quad \text{a.e. } t\in(0,T),\quad |a(t,u,v)|\leq M\|u\|\cdot\|v\|,$$

(6) $$\exists \alpha>0,\quad \forall v\in V,\quad \text{a.e. } t\in(0,T),\quad a(t,v,v)\geq\alpha\|v\|^2.$$

With $a(t,\cdot,\cdot)$ and its adjoint $a^*(t,u,v)=a(t,v,u)$ are associated two continuous operators from V into V' defined by

$$\forall u,v\in V,\quad (A(t)u,v)=a(t,u,v),\quad (A^*(t)u,v)=a^*(t,u,v).$$

We will also assume that $A(t)$ and $A^*(t)$ satisfy maximum principle properties, namely that the contractions $r\mapsto|r|$ and $r\mapsto r^+\wedge 1$ operate on V equipped with a and a^* that is

(7) $$\forall v\in V,\quad v^+\in V,\quad v^-\in V \text{ and a.e. } t\in(0,T),\quad a(t,v^+,v^-)\geq 0,$$

(8) $$\begin{aligned}&\forall v\in V,\quad v^+\wedge 1\in V \text{ and a.e. } t\in(0,T),\\ &a(t,u+u^+\wedge 1,u-u^+\wedge 1)\geq 0,\qquad a(t,u-u^+\wedge 1,u+u^+\wedge 1)\geq 0.\end{aligned}$$

Example 2. Corresponding to the choices of X and V in the examples above one can successively choose:

$$(\alpha)\quad a(t,u,v)=\sum_{i,j=1}^N\int_{\mathbb{R}^N}a_{ij}(x,t)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}dx+\sum_{i=1}^N\int_{\mathbb{R}^N}b_i(x,t)\frac{\partial u}{\partial x_i}v\,dx$$

$$+\int_{\mathbb{R}^n}c_0(x,t)uv\,dx,$$

where $a_{ij}, b_i, c_0\in L^\infty([0,T[\times\mathbb{R}^N)$ and satisfy

$$\exists\alpha>0,\quad \forall\xi\in\mathbb{R}^N,\quad \sum_{i,j=1}a_{ij}\xi_i\xi_j\geq\alpha\left(\sum_{i=1}^N\xi_i^2\right)\quad \text{a.e. on } Q.$$

Then, $a(\cdot,\cdot,\cdot)$ satisfies (4) and (5). It satisfies (7) and (8) if $c_0\geq 0$ and satisfies (6) if $c_0\geq A$ for A large enough. Since we will study parabolic properties, the latter point is not a restriction.

(β), (γ) One can choose $a(\cdot,\cdot,\cdot)$ as above where one replaces $\mathbb{R}^N$ by Ω.

(δ) Take a defined by

$$\text{a.e. } t\in(0,T),\quad \forall u,v\in\mathbb{R},\quad a(t,u,v)=a(t)uv,$$

where $a\in L^\infty(0,T)$, $a\geq 0$.

Parabolic potentials.

DEFINITION 1. We shall call a *parabolic potential* any element of

$$\mathcal{P}=\left\{u\in L^2(0,T;V)\cap L^\infty(0,T;L^2(X));\ \forall v\in\mathcal{W} \text{ with } v(T)=0,\ v\geq 0,\right.$$
$$\left.\int_0^T\left[\left(-\frac{\partial v}{\partial t}(t),u(t)\right)+a(t,u(t),v(t))\right]dt\geq 0\right\}.$$

Remark. We will often omit the variable t in the integral above and write it as

$$\int_0^T\left(-\frac{\partial v}{\partial t},u\right)+a(u,v).$$

Thanks to the Hahn–Banach theorem, we have (see [8], [10]):

PROPOSITION 1. *Let $u\in\mathcal{P}$. Then there exists a unique Radon measure on Q, denoted by $\mathcal{E}u$, such that*

$$\forall v\in\mathcal{W}\cap\mathcal{K}(Q)\quad \textit{with } v(T)=0,\quad \int_0^T\left(-\frac{\partial v}{\partial t},u\right)+a(u,v)=\int_Q v\,d(\mathcal{E}u).$$

Details are given in [8], [10] about the space $\mathcal{P}$ and the measures $\mathcal{E}u$. Let us just make them explicit in a particular but typical example.

Example 3. Let $X=\Omega$, $V=H_0^1(\Omega)$, $V'=H^{-1}(\Omega)$ and

$$\forall t\in[0,T],\quad \forall u,v\in V,\quad a(t,u,v)=\int_\Omega\nabla u\nabla v.$$

Then, if $u\in L^2(0,T;H_0^1(\Omega))\cap L^\infty(0,T;L^2(\Omega))$,

$$(u\in\mathcal{P})\leftrightarrow\left(u\geq 0,\ \frac{\partial u}{\partial t}-\Delta u\geq 0 \text{ in } \mathcal{D}'(]0,T[\times\Omega)\right).$$

Moreover,

$$\mathcal{E}u=u(0^+)\,dx_0+\frac{\partial u}{\partial t}-\Delta u,$$

where dx_0 is the Lebesgue measure induced on $\{0\}\times\Omega$ and $u(0^+)=\operatorname{ess\,lim}_{t\to 0^+}u(t)$ in $L^2(\Omega)$.

More examples are given in [8].

2. The main result. Let us first recall the usual definition of the parabolic capacity associated with the operators $A(t)$.

For any *open* set ω of Q, we consider

$$\mathcal{P}_\omega=\{u\in\mathcal{P};\ u\geq 1 \text{ a.e. on } \omega\}.$$

Then, if $\mathcal{P}_\omega$ is not empty, it has a smallest element u_ω called the *capacitary potential* of ω (see [8], [10] for a proof).

DEFINITION 2. For any *open* set $\omega \subset Q$, we set

$$c_0(\omega) = \begin{cases} \int_Q d\mathfrak{E} u_\omega & \text{if } \mathcal{P}_\omega \neq \varnothing, \\ +\infty & \text{if } \mathcal{P}_\omega = \varnothing. \end{cases}$$

For any $E \subset Q$, we define:

$$\text{capacity of } E = c_0(E) = \inf_{\substack{\omega \supset E \\ \omega \text{ open}}} c_0(\omega).$$

Now let us define two different capacities. For that we denote by Λ the space $\mathcal{V} \cap L^\infty(0,T; L^2(X))$ with the norm:

$$\|u\|_\Lambda^2 = \|u\|_{\mathcal{V}}^2 + \sup_{t \in (0,T)} \operatorname{ess} |u(t)|_2^2.$$

DEFINITIONS 3 AND 4. For any *open* set $\omega \subset Q$, we set:

$$c_1(\omega) = \inf\{\|u\|_\Lambda;\ u \in \mathcal{P},\ u \geq 1 \text{ a.e. on } \omega\},$$

$$c_2(\omega) = \inf\{\|v\|_{\mathcal{W}};\ v \in \mathcal{W}^+,\ v \geq 1 \text{ a.e. on } \omega\}.$$

For any $E \subset Q$, we define:

$$c_1(E) = \inf_{\substack{\omega \supset E \\ \omega \text{ open}}} c_1(\omega), \qquad c_2(E) = \inf_{\substack{\omega \supset E \\ \omega \text{ open}}} c_2(\omega).$$

Then, we have the main result.

THEOREM 1. *There exist* $a, b > 0$ *such that, for any* $E \subset Q$:

(i) $a \cdot c_0(E) \leq [c_1(E)]^2 \leq b \cdot c_0(E)$,

(ii) $a \cdot c_0(E) \leq [c_2(E)]^2 \leq b \cdot c_0(E)$.

Remarks. According to this result, to estimate the parabolic capacity of a set E, one can

(i) find $u \in \mathcal{P}$ with $u \geq 1$ on a neighborhood of E and compute the Λ-norm of u, or

(ii) find $v \in \mathcal{W}$ with $v \geq 1$ on a neighborhood of E and compute the $\mathcal{W}$-norm of v.

Note that the definition of $c_1(\cdot)$ still involves $\mathcal{P}$ and hence the operators $A(t)$, but it uses the Hilbert norms of $\mathcal{V}$ and $L^2(X)$ instead of an "L^1-norm" as in the definition of $c_0(\omega)$.

The interest of the definition of $c_2(\cdot)$ is that it only involves the topology of the Hilbert space $\mathcal{W}$ and does not depend on the operators $A(t)$.

Recall that $\mathcal{W} \hookrightarrow \Lambda$; so the topology of Λ is weaker than the topology of $\mathcal{W}$. But it is also sufficient to estimate the capacity of a set if one uses elements of $\mathcal{P}$.

If $c_1(\cdot)$ and $c_2(\cdot)$ are not generally "strong" capacities, they are however "weak" capacities. Namely:

PROPOSITION 2. (i) *For* $i = 0, 1, 2$:

(a) $E_1 \subset E_2 \Rightarrow c_i(E_1) \leq c_i(E_2)$.

(b) *For any nondecreasing sequence* (E_n) *of subsets of* Q

$$c_i\Big(\bigcup_n E_n\Big) = \sup_n c_i(E_n).$$

(c) *For any nonincreasing sequence* (K_n) *of compacts of* Q

$$c_i\Big(\bigcap_n K_n\Big) = \inf_n c_i(K_n).$$

(ii) (*strong subadditivity*). *For all* $E_1, E_2 \subset Q$,

$$c_0(E_1 \cup E_2) + c_0(E_1 \cap E_2) \leq c_0(E_1) + c_0(E_2).$$

(iii) ("*weak*" *subadditivity*). *For* $i = 1, 2$, *for all* $E_1, E_2 \subset Q$,

$$c_i(E_1 \cup E_2) \leq c_i(E_1) + c_i(E_2).$$

The properties of $c_0(\cdot)$ have already been studied in [8] (or [10]); we shall not reproduce the proofs here.

Only the property (b) is difficult for $c_1(\cdot)$ and $c_2(\cdot)$. It will result from important properties of the spaces $\mathcal{P}$ and $\mathcal{W}$ that will also be used to prove the part (ii) of Theorem 1. But let us begin by the proof of (i) in Theorem 1 which is fairly easy.

Proof of (i) *in Theorem* 1. It is sufficient to prove it for any open set $\omega \subset Q$.

Let us prove that, if $\mathcal{P}_\omega \neq \varnothing$,

$$\|u_\omega\|_\Lambda^2 \leq \left(2 + \frac{1}{\alpha}\right) c_0(\omega). \tag{9}$$

In order to compute, we need to approximate u_ω by more "regular" potentials. This is the purpose of the [8, Thm. I-1] (see also [9]) which says that the solution of:

$$u_\lambda \in \mathcal{W}, \quad u_\lambda(0) = u_\omega(0), \quad u_\lambda + \lambda\left(\frac{\partial u_\lambda}{\partial t} + Au_\lambda\right) = u_\omega \quad (\lambda > 0) \tag{10}$$

satisfies

$$u_\lambda \in \mathcal{P}, \quad u_\lambda \leq u_\omega, \quad \int_Q d\mathcal{E}u_\lambda \leq \int_Q d\mathcal{E}u_\omega,$$

and converges in $L^2(0,T;L^2(X))$ and weakly in $\mathcal{V}$ to u_ω when $\lambda \to 0^+$. But for any $t \in (0,T)$

$$\frac{1}{2}|u_\lambda(t)|_2^2 + \frac{1}{2}|u_\lambda(0)|_2^2 + \int_0^t a(u_\lambda, u_\lambda) = \int_0^t \left(\frac{\partial u_\lambda}{dt} + Au_\lambda, u_\lambda\right) + (u_\lambda(0), \quad u_\lambda(0)).$$

Since $0 \leq u_\lambda \leq u_\omega \leq 1$, the right-hand side (which is formally equal to $\int_{[0,t]\times X} u_\lambda d\mathcal{E}u_\lambda$) is less than $\int_Q d\mathcal{E}u_\lambda$ (see [8, Prop. I-3]). Hence, for any λ, by (6)

$$\frac{1}{2}|u_\lambda(t)|_2^2, \ \alpha\|u_\lambda\|_{\mathcal{V}}^2 \leq \int_Q d\mathcal{E}u_\lambda \leq c_0(\omega).$$

Letting λ go to 0 gives (9) and the second inequality of (i) with $b = 2 + \frac{1}{\alpha}$.

For the first inequality, let $\omega \subset Q$ open and $u \in \mathcal{P}$ with $u \geq 1$ a.e. on ω. For any compact $K \subset \omega$, there exists $\psi \in \mathcal{K}(Q) \cap \mathcal{W}^+$ equal to 1 on K and with support in ω (see [8, Lemma II-2]). Then, if u_K is the capacitary potential of K, $\mathcal{E}u_K$ is carried by K (see [8], [10]). Therefore,

$$c_0(K) = \int_Q d\mathcal{E}u_K \leq \int_Q \psi d\mathcal{E}u_K. \tag{11}$$

Now, if u_λ is the solution of (10), where u_ω is replaced by u_K, since $\partial u_\lambda/\partial t + Au_\lambda \geq 0$ and $\psi \leq u$, we have

$$\int_Q \psi \, d\mathfrak{E} u_\lambda = (\psi(0), u_\lambda(0)) + \int_0^T \left(\frac{\partial u_\lambda}{\partial t} + Au_\lambda, \psi \right)$$

$$\leq (u(0), u_\lambda(0)) + \int_0^T \left(\frac{\partial u_\lambda}{\partial t} + Au_\lambda, u \right).$$

Using $u \in \mathcal{P}$, we obtain:

$$\int_Q \psi \, d\mathfrak{E} u_\lambda \leq (u(0), u_\lambda(0)) + (u(T), u_\lambda(T)) + \int_0^T a(u, u_\lambda) + a(u_\lambda, u).$$

When λ goes to 0^+, $\mathfrak{E} u_\lambda$ converges to $\mathfrak{E} u$ in the sense of measure. Hence, using (11), we have

$$c_0(K) \leq |u(0)|_2 |u_K(0)|_2 + |u(T)|_2 |u_K(T)|_2 + \int_0^T a(u, u_K) + a(u_K, u). \tag{12}$$

But if $\mathcal{P}_\omega \neq \varnothing$ there exists a nondecreasing sequence of compacts $K_n \subset \omega$ such that $c_0(K_n)$ converges to $c_0(\omega)$ and u_{K_n} weakly converges to u_ω in $\mathcal{V}$ (see, for instance, [8, Prop. II-4]). Then, passing to the limit in (12), we obtain that there exists c depending only on M (see (5)) such that:

$$c_0(\omega) \leq c \|u\|_\Lambda \|u_\omega\|_\Lambda .$$

This together with (9) completes the proof of (i) in Theorem 1.

Proof of (ii) *in Theorem* 1. It is a direct consequence of the part (i) and the following proposition.

PROPOSITION 3. *There exists* $k > 0$ *such that*

(i) $\forall u \in \mathcal{P}$, $\exists v \in \mathcal{W}$ with

$$v \geq u, \qquad \|v\|_{\mathcal{W}} \leq k \|u\|_\Lambda .$$

(ii) $\forall v \in \mathcal{W}$, $\exists u \in \mathcal{P}$ with

$$u \geq v^+, \qquad \|u\|_\Lambda \leq k \|v\|_{\mathcal{W}}.$$

Proof of Proposition 3. For (i), given $u \in \mathcal{P}$, we consider the solution v of

$$v \in \mathcal{W}, \quad v(T) = u(T^-), \quad -\frac{\partial v}{\partial t} + A^*(t)v = A^*(t)u + A(t)u. \tag{13}$$

By well-known results about these linear parabolic equations (see Lions–Magenes [5]), such a solution exists in $\mathcal{W}$ and there exists a constant c depending only on $A(t)$ such that

$$\|v\|_{\mathcal{W}} \leq c \left[|u(T)|_2 + \|A^*(t)u\|_{\mathcal{V}'} + \|A(t)u\|_{\mathcal{V}'} \right].$$

That is

$$\|v\|_{\mathcal{W}} \leq k \|u\|_\Lambda ,$$

where k depends only on $A(t)$. Moreover, we *formally* have:

$$-\frac{\partial}{\partial t}(v-u) + A^*(t)(v-u) = \frac{\partial u}{\partial t} + A(t)u \geq 0 \quad (\text{since } u \in \mathcal{P}).$$

Since $(v-u)(T)=0$, by the maximum principle, $v\geq u$. This formal computation can be justified in the following way. Given $f\in L^2(0,T;L^2(X))$, $f\geq 0$, let us consider the solution w of:

$$w\in\mathcal{W},\quad w(0)=0,\quad \frac{\partial w}{\partial t}+A(t)w=f.$$

By the maximum principle (see (7)), $f\geq 0\Rightarrow w\geq 0$. But

$$\int_0^T\left(\frac{\partial w}{\partial t}+A(t)w,v\right)=(v(T),w(T))+\int_0^T\left(-\frac{\partial v}{\partial t}+A^*(t)v,w\right).$$

This implies

$$\int_0^T(f,v-u)=(u(T),w(T))+\int_0^T\left(-\frac{\partial w}{\partial t},u\right)+a(u,w).$$

Since $w\geq 0$ and $u\in\mathcal{P}$, the right-hand side is nonnegative. As f is arbitrary, this implies $v\geq u$.

For (ii), given $v\in\mathcal{W}$, we consider

$$u=\inf\{w\in\mathcal{P};\ w\geq v\}=\inf\{w\in\mathcal{P};\ w\geq v^+\}. \tag{14}$$

Using the results of Mignot–Puel [6], it can be shown (see also [8, Lemma II-1]) that $u\in\mathcal{P}$ and is the limit in $L^2(0,T;L^2(X))$ and weakly in $\mathcal{V}$ of the solution u_ε of the penalized problem

$$u_\varepsilon\in\mathcal{P},\quad u_\varepsilon(0)=v(0),\quad \frac{\partial u_\varepsilon}{\partial t}+A(t)u_\varepsilon-\frac{1}{\varepsilon}(u_\varepsilon-v)^-=0\qquad(\varepsilon>0).$$

But, for any $t\in(0,T)$,

$$\begin{aligned}\frac{1}{2}|u_\varepsilon(t)|_2^2-\frac{1}{2}|v(0)|_2^2+\int_0^t a(u_\varepsilon,u_\varepsilon)&=\int_0^t\left(\frac{\partial u_\varepsilon}{\partial t}+Au_\varepsilon,u_\varepsilon\right)\\&=\int_0^t\left(\frac{1}{\varepsilon}(u_\varepsilon-v)^-,u_\varepsilon-v\right)+\int_0^t\left(\frac{\partial u_\varepsilon}{\partial t}+Au_\varepsilon,v\right)\\&\leq(u_\varepsilon(t),v(t))-(v(0),v(0))+\int_0^t\left(-\frac{\partial v}{\partial t}+A^*v,u_\varepsilon\right).\end{aligned}$$

Passing to the limit gives

$$\frac{1}{2}|u(t)|_2^2+\alpha\|u\|_{\mathcal{V}}^2\leq|u(t)|_2|v(t)|_2+\left\|-\frac{\partial v}{\partial t}+A^*v\right\|_{\mathcal{V}'}\cdot\|u\|_{\mathcal{V}}.$$

Hence, there exists a constant k depending only on $A(t)$ such that:

$$\|u\|_\Lambda^2\leq k\|v\|_{\mathcal{W}}\cdot\|u\|_\Lambda.$$

Since $u\in\mathcal{P}$ and $u\geq v^+$, this completes the proof.

In order to prove Proposition 2, let us introduce for any $E\subset Q$:

$$\mathcal{W}_E=\left\{v\in\mathcal{W}^+;\ v=\lim_{n\to\infty}v_n \text{ in } \mathcal{W} \text{ with } v_n\geq 1 \text{ a.e. on a neighborhood of } E\right\},$$

$$\mathcal{P}_E=\left\{u\in\mathcal{P};\ \exists\, u_n\in\mathcal{P} \text{ with } u=\lim u_n \text{ in } \mathcal{V},\ \limsup_{n\to\infty}\|u_n\|_\Lambda\leq\|u\|_\Lambda\right\},$$

$$u(T)=\lim_{n\to\infty}\{u_n(T) \text{ in } L^2(X) \text{ and } u_n\geq 1 \text{ on a neighborhood of } E\}.$$

If $E=\omega$ is an open set, we immediately have:

$$\mathcal{W}_\omega=\{v\in\mathcal{W}^+;\ v\geq 1 \text{ a.e. on } \omega\},\qquad \mathcal{P}_\omega=\{u\in\mathcal{P};\ u\geq 1 \text{ a.e. on } \omega\}.$$

Moreover, we verify that, for any $E\subset Q$:

$$c_1(E)=\inf\{\|u\|_\Lambda;\ u\in\mathcal{P}_E\},\qquad c_2(E)=\inf\{\|v\|_{\mathcal{W}};\ v\in\mathcal{W}_E\}.$$

Remark that $\mathcal{W}_E$ is a closed convex set in $\mathcal{W}$. Hence, if v_E is the projection of 0 on $\mathcal{W}_E$ in a Hilbert space $\mathcal{W}$, then $c_2(E)=\|v_E\|_{\mathcal{W}}$.

LEMMA 1. *For any nondecreasing sequence (E_n) of subsets of Q:*

(i) $\bigcap_n \mathcal{W}_{E_n}=\mathcal{W}_{\cup_n E_n}$,

(ii) $\bigcap_n \mathcal{P}_{E_n}=\mathcal{P}_{\cup_n E_n}$.

To prove Lemma 1, we will need the following consequence of Proposition 3.

LEMMA 2. *There exists $k>0$ such that, for any $v\in\mathcal{W}$, there exists $w\in\mathcal{W}$ with*

$$w\geq|v|,\qquad \|w\|_{\mathcal{W}}\leq k\|v\|_{\mathcal{W}}.$$

Proof of Lemma 2. Let $v\in\mathcal{W}$, by (ii) in Proposition 3, there exist $u_1,u_2\in\mathcal{P}$ such that

$$u_1\geq v^+,\quad u_2\geq v^-,\quad \|u_1\|_\Lambda,\|u_2\|_\Lambda\leq k\|v\|_{\mathcal{W}}.$$

Now by (i) of the same proposition, there exists $w\in\mathcal{W}$ with

$$w\geq u_1+u_2,\qquad \|w\|_{\mathcal{W}}\leq k\|u_1+u_2\|_\Lambda.$$

Then, $w\geq v^+ + v^- = |v|$ and satisfied

$$\|w\|_{\mathcal{W}}\leq 2k^2\|v\|_{\mathcal{W}}.$$

Remark. As a consequence of (7), if $v\in V$, then v^+, v^- and $|v|$ also belong to V and the norm of $|v|$ in V can be estimated in terms of the norm of v.

But there is no such estimate in $\mathcal{W}$ (see L. Tartar's remark in the appendix). However Lemma 2 will be sufficient for our purpose.

Proof of Lemma 1. Let $E=\cup_n E_n$; the inclusions $\mathcal{W}_E\subset\bigcap_n\mathcal{W}_{E_n}$, $\mathcal{P}_E\subset\bigcap_n\mathcal{P}_{E_n}$ are obvious.

Let $v\in\bigcap_n\mathcal{W}_{E_n}$; then there exists $v_n\in\mathcal{W}$ with $v_n\geq 1$ on a neighborhood ω_n of E_n and $\|v-v_n\|_{\mathcal{W}}\leq 2^{-n}$. The series $\Sigma_1^\infty(v_{n+1}-v_n)$ is converging in $\mathcal{W}$. By Lemma 2, there exists $w_n\in\mathcal{W}$ with

$$w_n\geq|v_{n+1}-v_n|,\qquad \|w_n\|_{\mathcal{W}}\leq k\|v_{n+1}-v_n\|_{\mathcal{W}}.$$

Hence the series $\Sigma_1^\infty w_n$ converges in $\mathcal{W}$.

Now set $g_n=v_n+\Sigma_n^\infty w_k$. If $k>n$

$$g_n\geq v_n+\sum_n^{k-1} w_j\geq v_n+\sum_n^{k-1}(v_{j+1}-v_j)=v_k\geq 1\quad \text{a.e. on } \omega_k.$$

Hence, $g_n\geq 1$ almost everywhere on $\cup_{n+1}^\infty\omega_k$ which is a neighborhood of E and $v=\lim g_n$ in $\mathcal{W}$. Therefore $v\in\mathcal{W}_E$.

Now let $u\in\bigcap_n\mathcal{P}_{E_n}$; then there exists $u_n\in\mathcal{P}$ such that $\|u-u_n\|_V+|u(T)-u_n(T)|_2\leq\frac{1}{n}$ and $u_n\geq 1$ on a neighborhood ω_n of E_n. For any $\lambda>0$, we consider the solution of

$$v_n^\lambda\in\mathcal{W},\quad v_n^\lambda(T)=u_n(T),\quad v_n^\lambda+\lambda\left(-\frac{\partial v_n^\lambda}{\partial t}+A^*v_n^\lambda\right)=u_n+\lambda(Au_n+A^*u_n).$$

Then, by [8, Lemma IV-1], $v_n^\lambda \geq u_n$. (Remark that formally $v_n^\lambda - u_n + \lambda(-\frac{\partial}{\partial t}(v_n^\lambda - u_n) + A^*(v_n^\lambda - u_n)) = \lambda(\frac{\partial u_n}{\partial t} + Au_n) \geq 0$.) Moreover, for λ fixed, v_n^λ converges in $\mathcal{W}$ to the solution of

$$v^\lambda \in \mathcal{W}, \quad v^\lambda(T) = u(T), \quad v^\lambda + \lambda\left(-\frac{\partial v^\lambda}{\partial t} + A^* v^\lambda\right) = u + \lambda(Au + A^*u).$$

Indeed,

$$\|v_n^\lambda - v^\lambda\|_{\mathcal{W}} \leq c_\lambda\left(\|u_n - u\|_{\mathcal{V}} + |u_n(T) - u(T)|_2\right).$$

Since $v_n^\lambda \geq u_n \geq 1$ on ω_n, as in the proof of Lemma 2, for any $\lambda > 0$, we can construct $g_n^\lambda \in \mathcal{W}$ converging in $\mathcal{W}$ to v^λ with $g_n^\lambda \geq 1$ on a neighborhood of E. Let us choose $g_\lambda = g_{n_\lambda}^\lambda$ such that $\|g_\lambda - v^\lambda\|_{\mathcal{W}} \leq \lambda$.

By Proposition 3, there exists $u_\lambda \in \mathcal{P}$ with $u_\lambda \geq g_\lambda - v^\lambda$ and $\|u_\lambda\|_\Lambda \leq k\|g_\lambda - v^\lambda\|_{\mathcal{W}} \leq k\lambda$. Moreover, by the results in [8, §IV], there exists a convex combination of the v^λ (still denoted by v^λ) such that:

- v^λ converges to u in $\mathcal{V}$,
- $\lim_{\lambda \to \infty^+} \|v^\lambda\|_\Lambda = \|u\|_\Lambda$,
- if $\hat{u}_\lambda = \inf\{u \in \mathcal{P};\ u \geq v^\lambda\}$, $\hat{u}_\lambda - v^\lambda$ converges to 0 in Λ.

Then $u_\lambda + \hat{u}_\lambda \in \mathcal{P}$, $u_\lambda + \hat{u}_\lambda \geq g_\lambda \geq 1$ on a neighborhood of E, $u_\lambda + \hat{u}_\lambda$ converges to u in $\mathcal{V}$, $u_\lambda(T) + \hat{u}_\lambda(T)$ converges to $u(T)$ in $L^2(X)$ and $\lim_{\lambda \to 0} \|u_\lambda + \hat{u}_\lambda\|_\Lambda = \|u\|_\Lambda$. Hence $u \in \mathcal{P}_E$.

Proof of Proposition 2. The properties of $c_0(\cdot)$ are shown in [8]. The part (a) of (i) is obvious. The point (b) is a direct consequence of Lemma 1.

For (c), remark that, for $i = 1, 2$,

$$c_i\left(\bigcap K_n\right) \leq \inf_n c_i(K_n).$$

Now, for $\varepsilon > 0$, there exists a neighborhood ω_ε of $K = \bigcap K_n$ such that

$$c_1(\omega_\varepsilon) \leq c_1(K) + \varepsilon, \qquad c_2(\omega_\varepsilon) \leq c_2(K) + \varepsilon.$$

But as K_n is a sequence of compacts decreasing to K, for n large enough, $K_n \subset \omega_\varepsilon$. Hence

$$\inf_n c_i(K_n) \leq c_i(K_n) \leq c_i(\omega_\varepsilon) \leq c_i(K) + \varepsilon.$$

For (iii), we use the subadditivity of $\|\cdot\|_{\mathcal{W}}$ and $\|\cdot\|_\Lambda$.

3. Application. We proved in [8] that the elements of $\mathcal{W}$ are quasicontinuous. We will give here a more direct proof using essentially the equivalent definition of the capacity given by Theorem 1 in terms of the $\mathcal{W}$-norm, together with Lemma 2. (See also [7] for abstract "elliptic" results of this kind.)

We recall that, given a capacity $c(\cdot)$ on Q:

DEFINITION. A function $v: Q \to \mathbb{R}$ is said to be *quasicontinuous* if there exists a nonincreasing sequence of open set $\omega_n \subset Q$ with

(i) $\lim_{n \to \infty} c(\omega_n) = 0$,

(ii) the restriction of v to the complement of ω_n is continuous for all n.

Remark. This definition is clearly invariant when one replaces $c(\cdot)$ by an "equivalent" capacity $\hat{c}(\cdot)$, that is a capacity satisfying for some $\alpha > 0$

$$\exists a, b > 0, \quad E \subset Q, \quad a \cdot c(E) \leq [\hat{c}(E)]^\alpha \leq b \cdot c(E).$$

Hence, the notion of quasicontinuity is the same for our capacities $c_0(\cdot)$, $c_1(\cdot)$ and $c_2(\cdot)$.

THEOREM 2. *Any element v of $\mathcal{W}$ has a unique quasicontinuous representation $\tilde{v}$.*

Remark. "Unique" means here that, if $\hat{v}$ is quasicontinuous and satisfies $\hat{v}=\tilde{v}$ almost everywhere, then $\hat{v}=\tilde{v}$ *quasieverywhere* (i.e., everywhere except on a set of zero capacity).

Proof of Theorem 2. Let $v\in\mathcal{W}$; by density of $\mathcal{K}(\tilde{Q})\cap\mathcal{W}$ in $\mathcal{W}$, there exist $v_n\in\mathcal{W}\cap\mathcal{K}(\tilde{Q})$ converging to v with

$$\sum_{n=1}^{\infty} 2^n\|v_{n+1}-v_n\|_{\mathcal{W}}<+\infty.$$

Let

$$\omega_n=\{z\in Q;\ |v_{n+1}(z)-v_n(z)|>2^{-n}\} \quad \text{and} \quad \Omega_p=\bigcup_{n\geq p}\omega_n.$$

By Lemma 2, there exists $w_n\in\mathcal{W}$ with

$$w_n\geq|v_{n+1}-v_n|, \qquad \|w_n\|_{\mathcal{W}}\leq k\cdot\|v_{n+1}-v_n\|_{\mathcal{W}}.$$

Hence

$$c_2(\omega_n)\leq c_2(\{z\in Q;\ w_n(z)>2^{-n}\})\leq 2^n\|w_n\|_{\mathcal{W}}.$$

This proves that $\lim_{p\to\infty} c_2(\Omega_p)=0$. But, for any p:

$$|v_{n+1}(z)-v_n(z)|\leq 2^{-n} \quad \forall z\notin\Omega_p, \quad \forall n\geq p.$$

Hence, v_n converges uniformly on the complement of each Ω_p. The limit $\tilde{v}$ is defined quasieverywhere (everywhere except on $\bigcap_p\Omega_p$ which is of zero capacity), $\tilde{v}$ is quasicontinuous and $\tilde{v}=v$ almost everywhere.

For the uniqueness, let us consider $\hat{v}$ quasicontinuous with $\tilde{v}=\hat{v}$ almost everywhere and ω_n a sequence of open sets associated with $\tilde{v}-\hat{v}$ (see the definition above). Then, $A_n=\{z\in Q;\ \tilde{v}-\hat{v}<0\}\cup\omega_n$ is open for any n. Since $\{z\in Q;\ \tilde{v}-\hat{v}<0\}$ is of measure 0, $\mathcal{W}_{A_n}=\mathcal{W}_{\omega_n}$ for all n. Hence

$$c_2\{z\in Q;\ \tilde{v}-\hat{v}<0\}\leq\lim_{n\to\infty} c_2(A_n)=\lim_{n\to\infty} c_2(\omega_n)=0.$$

Remark. The above property of the elements of $\mathcal{W}$ is a fundamental tool in the study of the structure of parabolic potentials as well as in the resolution of associated variational inequalities (see [9]).

Appendix (Communication of L. Tartar) (see Lemma 2).

PROPOSITION. *Given Ω a regular bounded set of $\mathbb{R}^N$, for $\mathcal{W}=\{v\in L^2(0,T;H_0^1(\Omega));\ \frac{\partial v}{\partial t}\in L^2(0,T;H^{-1}(\Omega))\}$, there does not exist any (continuous) function $C[\cdot]:[0,\infty)\to[0,\infty)$ such that*

$$\left\|\frac{\partial}{\partial t}|v|\right\|_{L^2(0,T;H^{-1})}\leq C\left[\|v\|_{\mathcal{W}}\right]. \tag{15}$$

Proof. Let $a\in H_0^1(\Omega)$ and $f_n\in W^{1,2}(0,1)$ with $f_n\geq 0$, $\|f_n'\|_{L^2(0,1)}=1$, f_n converges in $L^2(0,1)$ to 0 when n goes to ∞. (Take for instance $f_n(t)=\frac{\lambda}{n}[1+\sin n\pi t]$ with $\lambda=\sqrt{2}/\pi$).

Now, applying (15) to $v_n(t)=f_n(t)a$, since $|v_n|=f_n|a|$, one would have

$$\||a|\|_{H^{-1}(\Omega)}\leq C\left[\|f_n\|_{L^2(0,1)}\cdot\|a\|_{H_0^1}+\|a\|_{H^{-1}(\Omega)}\right].$$

That is,

$$\| |a| \|_{H^{-1}(\Omega)} \leq C\left[\|a\|_{H^{-1}(\Omega)}\right].$$

This is not true. (If $\Omega=(0,\pi)$ take, for instance, $a_n(x)=n\sin nx$.)

REFERENCES

[1] P. CHARRIER, *Contribution à l'étude de problèmes d'évolution*, Thèse, Univ. Bordeaux I, 1978.

[2] E. LANCONELLI, *Sul problema di Dirichlet per l'equazione del calore*, Ann. Mat. Pura ed Appl., 97 (1973), pp. 83–114.

[3] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[4] J.-L. LIONS AND G. STAMPACCHIA, *Variational inequalities*, Comm. Pure Appl., Math., 20 (1967), pp. 493–519.

[5] J.-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.

[6] F. MIGNOT AND J. P. PUEL, *Inéquations d'évolution paraboliques avec convexes dépendant du temps, Applications aux inéquations quasi-variationnelles d'évolution*, Arch. Rat. Mech. Anal., 64 (1977), pp. 59–91.

[7] H. ATTOUCH AND C. PICARD, *Problèmes variationnels et théorie du potentiel non linéaire*, Ann. Fac. Sci. Toulouse, 1, 1979, pp. 89–136.

[8] M. PIERRE, *Equations d'évolution non linéaires, inéquations variationnelles et potentiels paraboliques*. Thèse, Université Paris VI, 1979.

[9] _____, *Problèmes d'évolution avec contraintes unilatérales et potentiels paraboliques*, Comm. P.D.E., 4 (1979), pp. 1149–1197.

[10] _____, *Représentant précis d'un potentiel parabolique*, Sém. Th. du Potentiel, Univ. Paris VI, Lecture Notes in Mathematics 807, Springer, Berlin, 1980.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0011 $01.25/0

THE RIEMANN PROBLEM IN TWO SPACE DIMENSIONS FOR A SINGLE CONSERVATION LAW*

DAVID H. WAGNER†

Abstract. Solutions are given for the partial differential equation $\partial/\partial t\, u(t,x,y)+\partial/\partial x f(u(t,x,y))+\partial/\partial y\, g(u(t,x,y))=0$, with initial data constant in each quadrant of the (x,y) plane. This problem generalizes the Riemann problem for equations in one space dimension. Although existence and uniqueness of solutions are known, little is known concerning the qualitative behavior of solutions.

When f and g are convex and $f\equiv g$, then our solutions satisfy the uniqueness, or entropy condition given by Kruzkov and Vol'pert. Under certain extra conditions on f and g, our solutions satisfy the entropy condition if f and g are convex and sufficiently close. A counterexample is given to show the necessity of these extra conditions on f and g. The correct entropy solution for this counterexample exhibits new and interesting phenomena.

1. Introduction. Let f and g be given real functions satisfying $f''>0$ and $g''>0$. Consider the initial value problem

$$\frac{\partial}{\partial t}u(t,x,y)+\frac{\partial}{\partial x}f(u(t,x,y))+\frac{\partial}{\partial y}g(u(t,x,y))=0, \tag{1.1}$$

$$u(0,x,y)=\begin{cases} u_1 & \text{for } x>0, \quad y>0,\\ u_2 & \text{for } x<0, \quad y>0,\\ u_3 & \text{for } x<0, \quad y<0,\\ u_4 & \text{for } x>0, \quad y<0. \end{cases} \tag{1.2}$$

This is a Riemann problem in two space variables. It generalizes the Riemann problem in one space variable, the study of which has been a key to the understanding of solutions to systems of nonlinear hyperbolic conservation laws in one space variable [2].

Global existence of weak solutions to (1.1) with more general initial data than (1.2) was first proved by Conway and Smoller [1]. Later Vol'pert [9] and Kruzkov [6] proved existence and uniqueness of weak solutions satisfying an entropy condition, in the class of bounded measurable functions.

No similar advances have been made concerning systems of nonlinear hyperbolic conservation laws in two or more space variables; it may be that study of the Riemann problem for these systems will yield a breakthrough. In this paper we begin an attack on this problem by finding explicit entropy solutions to (1.1), (1.2) for a large class of pairs (f,g), in the case of a scalar conservation law.

We should mention that Guckenheimer [5] and Val'ka [8] have studied examples of (1.1) with piecewise constant initial data, in configurations different from (1.2).

DEFINITION 1.1. A bounded, measurable function $u:\mathbb{R}^+\times\mathbb{R}^2\to\mathbb{R}$ is said to be a weak solution to the initial value problem consisting of (1.1) with initial data $u(0,x,y)=u_0(x,y)$ if

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}^2} u\frac{\partial}{\partial t}\phi+f(u)\frac{\partial}{\partial x}\phi+g(u)\frac{\partial}{\partial y}\phi\, dx\, dy\, dt=0, \tag{1.3}$$

for every test function $\phi\in C_0^\infty(\mathbb{R}^+\times\mathbb{R}^2)$, and if $u(t,\cdot,\cdot)\to u_0$ in L^1_{loc} as $t\to 0$.

* Received by the editors April 7, 1982.

† Department of Mathematics, University of Houston, Houston, Texas 77004. This work was sponsored by the U.S. Army under contract DAAG 29-80-C-0041.

DEFINITION 1.2. (Vol'pert [8], Kruzkov [6]). A weak solution u, to (1.1), is said to satisfy the entropy condition if for any real constant k and any $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^2)$ such that $\phi \geq 0$,

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}^2} \operatorname{sign}(u-k)\Big[(u-k)\frac{\partial}{\partial t}\phi + (f(u)-f(k))\frac{\partial}{\partial x}\phi + (g(u)-g(k))\frac{\partial}{\partial y}\phi\Big]\,dx\,dy\,dt \geq 0. \tag{1.4}$$

We shall construct weak solutions to (1.1), (1.2) which are valid for any choice of the constants u_1, u_2, u_3, and u_4. It is conceivable that this construction may fail to produce a well defined function if f and g are not sufficiently close to each other. The degree of closeness that is sufficient may depend on the choice of f and g; therefore, in order to be rigorous we will state and prove our theorems in terms of the distance from f and g to a given reference function h. Thus we consider the pair (f,g) as a perturbation from the pair (h,h). Of course, our theorems will cover the case where f is held fixed and g is perturbed away from f.

The form of our solution to (1.1), (1.2) varies with different orderings of the constants u_i. Thus there are twenty-four cases to be considered. Fortunately, these twenty-four cases can be reduced, via geometrical reflections and inversions, to eight. We will however, consider each of the twenty-four cases and identify the reductions.

We shall show that our construction produces a well defined function which satisfies the entropy condition, if $f \equiv g$, and $f'' > 0$; see Theorem 1. We will also show in Theorem 2, that under certain ordering conditions on u_1, u_2, u_3 and u_4, we have that for every function h such that $h'' > 0$, there exists $\varepsilon > 0$ such that $\|f-h\|_{C^2} < \varepsilon$ and $\|g-h\|_{C^2} < \varepsilon$ imply that our construction produces a well defined function which satisfies the entropy condition. In Theorem 3 we shall show that if the ordering conditions of Theorem 2 are not satisfied, then, provided $f'' = g''$ and $f''' = g'''$ at certain points w to be specified later, for every h such that $h'' > 0$, and $h'' = f''$ and $h''' = f'''$ at all of the points w, there exists $\varepsilon > 0$ such that $\|f-h\|_{C^4} < \varepsilon$ and $\|g-h\|_{C^4} < \varepsilon$ imply that our construction produces a well defined function which satisfies the entropy condition.

Finally, an example will be given where $f''' \neq g'''$ at the point w of Theorem 3, and where our constructed solution, although it is a weak solution, does not satisfy the entropy condition. In this example f and g may be arbitrarily close, and the initial data may be arbitrarily small. We will also give the correct entropy solution for this example.

One system of equations containing a scalar conservation law is that describing two-phase, two-dimensional immiscible flow in porous media, where gravity, capillary pressure, molecular diffusion, compressibility, as well as spatial variations in porosity, depth and viscosity, have been neglected (see [3], [4], [7]):

$$\frac{\partial S}{\partial t} + \frac{\partial}{\partial x}(v_1 f(S)) + \frac{\partial}{\partial y}(v_2 f(S)) = 0, \tag{1.5}$$

$$\mathbf{v} = (v_1, v_2) = -k(S)\nabla p, \tag{1.6}$$

$$\nabla \cdot \mathbf{v} = \text{source terms}. \tag{1.7}$$

In [3], [4], $f(S) = s^2/k(S)$ was used. Although one may imagine that our solutions are thus special entropy solutions of this system with ∇p constant, (1.7) prevents this. However some of our solutions, Cases 9 and 19, exhibit shock waves meeting at an

acute angle, or in a cusp, similar to the fingering behavior which is of interest in oil recovery problems, and for which (1.5)–(1.7) is a model.

2. Construction of the solutions. We shall see that our constructed solutions are piecewise smooth, having discontinuity sets consisting almost everywhere with respect to two-dimensional Hausdorff measure, of smooth surfaces. In this context, Definition 1.2 implies two conditions, given below, on the discontinuities of a solution. These can easily be derived via localization and integration by parts and appropriate choices of the constant k.

Condition 2.1 (the Rankine–Hugoniot condition). At any point p on a surface of discontinuity S of the solution u, if

(a) $\mathbf{n}$ is a unit normal vector to S at p,

(b) $u^+ = \lim_{\varepsilon \to 0^+} u(p+\varepsilon \mathbf{n})$,

(c) $u^- = \lim_{\varepsilon \to 0^+} u(p-\varepsilon \mathbf{n})$,

then

$$\mathbf{n} \cdot (u^+ - u^-, f(u^+) - f(u^-), g(u^+) - g(u^-)) = 0. \tag{2.1}$$

Condition 2.2 (the entropy condition). Orient $\mathbf{n}$ so that $u^+ \geq u^-$. If k is any constant such that $u^- \leq k \leq u^+$, then

$$\mathbf{n} \cdot (k - u^+, f(k) - f(u^+), g(k) - g(u^+)) \geq 0. \tag{2.2}$$

Using (2.1) one may check that (2.2) is equivalent to

$$\mathbf{n} \cdot (k - u^-, f(k) - f(u^-), g(k) - g(u^-)) \geq 0. \tag{2.3}$$

One may further check that if a function u is a piecewise classical solution, except for smooth surfaces of discontinuity where Conditions 2.1 and 2.2 hold, then u is a weak solution satisfying the entropy condition.

Let us consider one-dimensional shock waves and rarefaction waves, as they arise in the two-dimensional Riemann problem.

(a) *The one-dimensional shock wave.* If the initial data is

$$u(0,x,y) = \begin{cases} u_1 & \text{if } x < 0, \\ u_2 & \text{if } x > 0, \end{cases} \tag{2.4}$$

and $u_1 > u_2$ then the problem really has only one space dimension:

$$u_t + f(u)_x = 0, \qquad u(0,x) = \begin{cases} u_1 & \text{if } x < 0, \\ u_2 & \text{if } x > 0. \end{cases} \tag{2.5}$$

In this case the solution is well known:

$$u(t,x,y) = u(t,x) = \begin{cases} u_1 & \text{if } x \leq \dfrac{f(u_1) - f(u_2)}{u_1 - u_2} t, \\ u_2 & \text{if } x \geq \dfrac{f(u_1) - f(u_2)}{u_1 - u_2} t. \end{cases} \tag{2.6}$$

This solution has a discontinuity, called a "shock wave", along the plane

$$x = \frac{f(u_1) - f(u_2)}{u_1 - u_2} t. \tag{2.7}$$

We will refer to this shock wave as $SX[u_1,u_2]$, for "shock in the x direction connecting u_1 to u_2." The shock wave in the y direction obtained by interchanging x with y and f with g above we will call $SY[u_1,u_2]$.

(b) *The one-dimensional rarefaction wave.* If the initial data is

$$u(0,x,y)=\begin{cases}u_1 & \text{if } x<0,\\ u_2 & \text{if } x>0,\end{cases} \tag{2.8}$$

and $u_1<u_2$, then the problem again reduces to (2.5), and its solution is well known:

$$u(t,x,y)=u(t,x)=\begin{cases}u_1 & \text{if } x<f'(u_1)t,\\ u_2 & \text{if } x>f'(u_2)t,\\ s & \text{if } x=g'(s)t,\quad u_1\le s\le u_2.\end{cases} \tag{2.9}$$

The part of this solution between $x=f'(u_1)t$ and $x=f'(u_2)$ is called a "rarefaction wave". We will refer to this particular rarefaction wave as $RX[u_1,u_2]$, for "rarefaction in the x-direction, connecting u_1 to u_2."

Note that the solution described in (a) is a weak solution to (b), since it satisfies Condition 2.1. However, it does not satisfy the entropy condition since $u_1<u_2$, and f is assumed to be convex.

The interaction of RX with SY. Let the initial data be as in (1.2), with $u_1=u_2=u_3=w$, and $u_4=v$, and $v>w$. Then for bounded t, and sufficiently large x, the solution looks locally like $SY[v,w]$, due to the principle of finite domain of dependence, which was shown to hold in this context by Vol'pert [9] and Kruzkov [6]. For y sufficiently negative, the solution looks locally like $RX[w,v]$. Since a solution is invariant under dilations $(t,x,y)\to(ct,cx,cy)$ for $c>0$ whenever the initial data $u(0,x,y)$ is invariant under dilations $(x,y)\to(cx,cy)$, we may describe a solution completely by describing it along the plane $t=1$. The solution is constant on rays through the (t,x,y) origin.

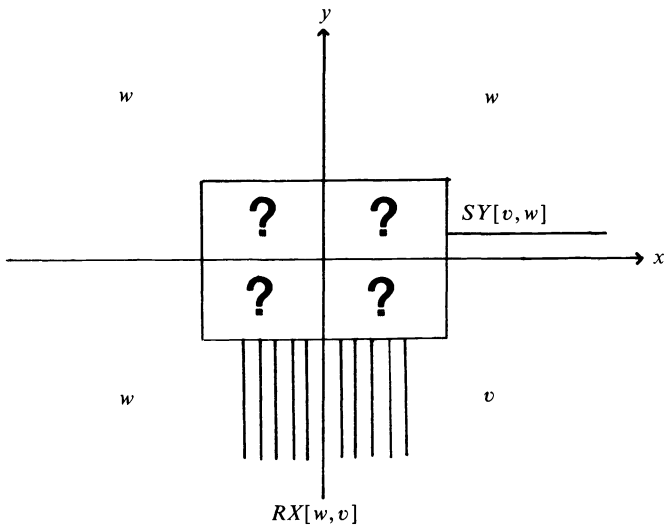


FIG. 1. *Interaction of RX with SY.*

Thus our current knowledge of the solution to this problem may be described by Fig. 1. In Fig. 1, the horizontal line labelled $SY[v,w]$ indicates the plane of that shock wave, and the vertical lines labelled $RX[w,v]$ indicate planes along which the solution u

is constant, thereby depicting a rarefaction wave. The space labelled ? is filled in as follows. The RX region meets the region where $u=w$ along a smooth surface of discontinuity S, having equations $x=f'(s)t$, $y=\gamma(s)t$, parametrized by s,t for $w\leq s\leq v$. The unknown function γ is determined by the jump conditions, as follows.

First we describe the normal vector $\mathbf{n}=(n_t,n_x,n_y)$, to the shock surface S in terms of f and γ. To do this, we need two tangent vectors to the surface $x=f'(s)t, y=\gamma(s)t$.

Holding s fixed, we have $dx=f'(s)\,dt$, $dy=\gamma(s)\,dt$. Holding t fixed, we have $dx=f''(s)t\,ds$, $dy=\gamma'(s)t\,ds$. Thus we have two tangent vectors, $\mathbf{v}_1=(1,f'(s),\gamma(s))$, $\mathbf{v}_2=(0,f''(s),\gamma'(s))$. Then

$$\mathbf{n}=\mathbf{v}_1\times\mathbf{v}_2=(-f''(s)\gamma(s)+f'(s)\gamma'(s),-\gamma'(s),f''(s)). \tag{2.10}$$

Keeping in mind that in $RX[w,v]$, $u=s$ on the plane $x=f'(s)t$ for $y<\gamma(s)t$, the Rankine–Hugoniot condition gives us

$$\begin{aligned}(w-s)(f'(s)\gamma'(s)-f''(s)\gamma(s))+(f(w)-f(s))(-\gamma'(s))\\ +f''(s)(g(w)-g(s))=0.\end{aligned} \tag{2.11}$$

This equation is a first order, linear, scalar differential equation for the unknown function γ:

$$\gamma'(s)=f''(s)\,\frac{g(w)-g(s)-\gamma(s)(w-s)}{f(w)-f(s)-f'(s)(w-s)}. \tag{2.12}$$

Note that the denominator in the right-hand side of (2.12) is always positive for $w\neq s$, since $f''>0$.

The shock surface S should be a smooth continuation of the planar shock wave $SY[v,w]$ through the rarefaction wave $RX[w,v]$. Therefore these two surfaces should meet along a common line at the right edge of $RX[w,v]$. This yields the following initial condition for γ:

$$\gamma(v)=\frac{g(v)-g(w)}{v-w}. \tag{2.13}$$

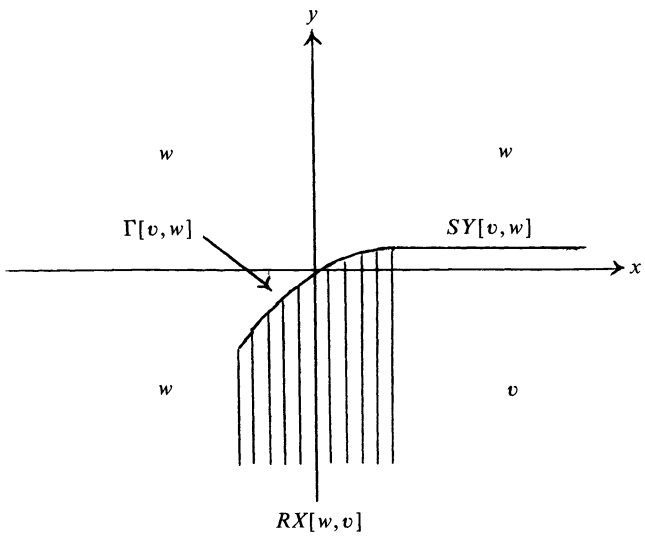


FIG. 2. *The shock surface* $\Gamma[v,w]$.

Since (2.12) is linear, it is thus explicitly solvable. The following formula for γ can be obtained from the usual one by an integration by parts:

$$\gamma(s)=\frac{g(s)-g(w)}{s-w}$$

$$(2.14)\qquad -\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)}{f(w)-f(z)-f'(z)(w-z)}\,dz\right)\frac{g(w)-g(r)-g'(r)(w-r)}{(w-r)^2}\,dr.$$

Note that for $w<s<v$, we have $\gamma(s)>(g(s)-g(w))/(s-w)$; hence, using (2.12), we have $\gamma'(s)>0$. Therefore, as s decreases to w, $\gamma(s)$ approaches some finite limit, which is greater than or equal to $g'(w)$. In fact this limit is $g'(w)$, as we shall see in §3.

We shall refer to the shock surface S as $\Gamma[v,w]$. See Fig. 2. This shock surface will occur in solutions to the Riemann problem, and usually in truncated form, denoted here $\Gamma[v,w:p]$, namely the portion of $\Gamma[v,w]$ corresponding to values of s greater than some number p and less than v. Reflecting $\Gamma[v,w]$ across the line $x=y$ by interchanging x with y and f with g, we obtain a similar shock wave, denoted $\Gamma R[v,w]$.

Note that, given any solution u to (1.1), with initial data $u_0(x,y)$, the function $\tilde{u}$ given by $\tilde{u}(t,x,y)=-u(t,-x,-y)$ is a solution to

$$\tilde{u}_t+f(-\tilde{u})_x+g(-\tilde{u})y=0$$

with initial data $\tilde{u}(0,x,y)=-u_0(-x,-y)$. Thus, if $u(t,x,y)=F(t,x,y,f,g,u_1,u_2,u_3,u_4)$ is a formula giving the solution to the Riemann problem in terms of f,g, and the initial data constants $u_1,\cdots,u_4$ for a given ordering of these constants, then the solution for the case given by interchanging u_1 with u_3 and u_2 with u_4 in the given ordering, and then reversing the order, is $u(t,x,y)=-F(t,-x,-y,\check{f},\check{g},-u_3,-u_4,-u_1,-u_2)$, where $\check{f}(s)=f(-s)$. Thus, for example, the solution for the case $u_1<u_2<u_3<u_4$ determines the solution for the case $u_3>u_4>u_1>u_2$. We call this procedure "inversion."

If we apply the inversion process to $\Gamma[v,w]$, we get an inverted Γ-shock, which we shall call $\Gamma I[v,w]$. See Fig. 3. We call the reflection of $\Gamma I[v,w]$ across the line $y=x$, $\Gamma IR[v,w]$.

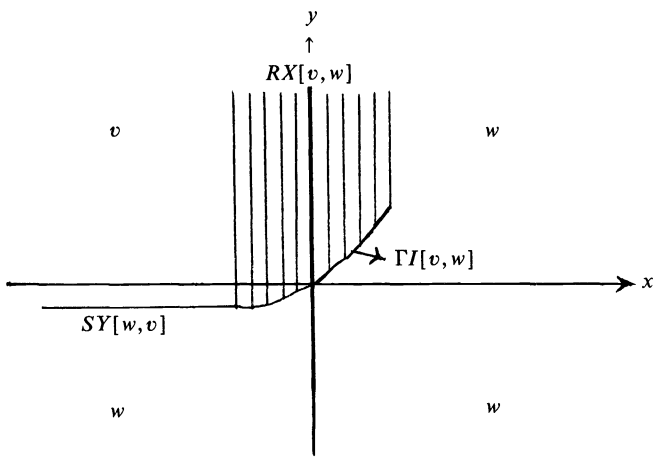


FIG. 3. *The shock surface* $\Gamma I[v,w]$.

We are now ready to discuss the Riemann problem.

The formula for the solution to the Riemann problem is different for different orderings of the initial data constants u_1, u_2, u_3 and u_4. This indicates that twenty-four cases must be considered; however reflections, inversions, and reflected inversions of cases previously discussed will only be indicated as such, and will not be discussed in detail. This reduces the number of cases to be discussed, to eight. The formula to be given consists of a picture for each case, with labels such as $RX[u_1,u_2]$, $SY[u_2,u_3]$, and $\Gamma[u_1,u_2]$ given to parts of the picture. The interested reader may then refer to the formula given earlier in this chapter for each of these phenomena.

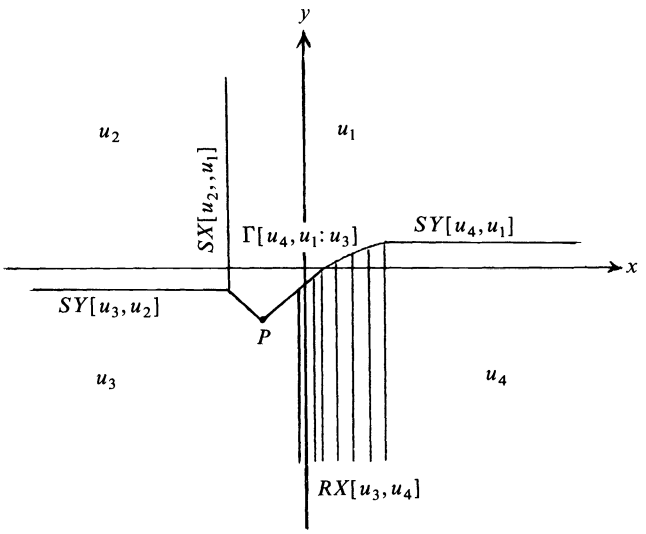


FIG. 4. *The case* $u_1<u_2<u_3<u_4$.

Case 1. $u_1<u_2<u_3<u_4$. At $t=1$ the solution looks like Fig. 4. In this case we have two shock waves extending from the intersection of $SY[u_3,u_2]$ and $SX[u_2,u_1]$ to the point P, and from the line $x=f'(u_3)t$, $y=\gamma(u_3)t$ to P. The following theorem concerning these shock waves is due to Guckenheimer [5].

THEOREM. *A shock surface between two constant states* a,b, *in a solution to the Riemann problem lies in a plane which passes through the line*

$$x=\frac{f(a)-f(b)}{a-b}t, \qquad y=\frac{g(a)-g(b)}{a-b}t. \tag{2.15}$$

Thus the two shock waves extending to P are planar. We shall call these shock waves "Ψ-shocks."

Now from (2.15) we deduce that P has coordinates

$$\left(\frac{f(u_3)-f(u_1)}{u_3-u_1}, \frac{g(u_3)-g(u_1)}{u_3-u_1}\right).$$

Note that since $u_1<u_2<u_3$, we have

$$\frac{f(u_2)-f(u_1)}{u_2-u_1}<\frac{f(u_3)-f(u_1)}{u_3-u_1}<f'(u_3); \tag{2.16}$$

these are the x-coordinates of $SX[u_2,u_1]$, P, and the left edge of $RX[u_3,u_4]$, respectively. Thus both Ψ-shocks may be parameterized by s and t with the equations

$x=f'(s)t$ and $y=\psi(s)t$. In case $u_1=u_2$ or $u_2=u_3$ then either the left or right Ψ-shock, respectively, is not present, because both end points of that particular shock are identical.

Using the method by which (2.12) was derived, one may derive the following differential equation for ψ:

$$\psi'(s)=f''(s)\,\frac{g(u_1)-g(u_3)-\psi(s)(u_1-u_3)}{f(u_1)-f(u_3)-f'(s)(u_1-u_3)}. \tag{2.17}$$

Note, from Fig. 4, that the right Ψ-shock meets the left endpoint of the $\Gamma[u_4,u_1:u_3]$ shock wave continuously, so that $\psi(u_3)=\gamma(u_3)$. Comparing (2.17) with (2.12), we see that $\psi'(u_3)=\gamma'(u_3)$. Thus the Γ-shock meets the Ψ-shock with first-order smoothness. We may also deduce that for any point $(f'(v),\gamma(v))$ on the $\Gamma[u_4,u_1]$ shock at $t=1$, the tangent line to the curve $x=f'(s)$, $y=\gamma(s)$ passes through the point P given by

$$P=\left(\frac{f(v)-f(u_1)}{v-u_1},\ \frac{g(v)-g(u_1)}{v-u_1}\right). \tag{2.18}$$

One may also deduce this by rewriting (2.12):

$$\frac{dy}{dx}=\frac{\gamma'(s)}{f''(s)}=\frac{\dfrac{g(u_1)-g(s)}{u_1-s}-\gamma(s)}{\dfrac{f(u_1)-f(s)}{u_1-s}-f'(s)}. \tag{2.19}$$

Case 2. $u_1<u_4<u_3<u_2$. This is the reflection across the line $y=x$ of Case 1.
Case 3. $u_3>u_4>u_1>u_2$. This is the inversion of Case 1.
Case 4. $u_3>u_2>u_1>u_4$. This is the reflected inversion of Case 1.
Case 5. $u_2<u_1<u_3<u_4$. At $t=1$ the solution looks like Fig. 5.

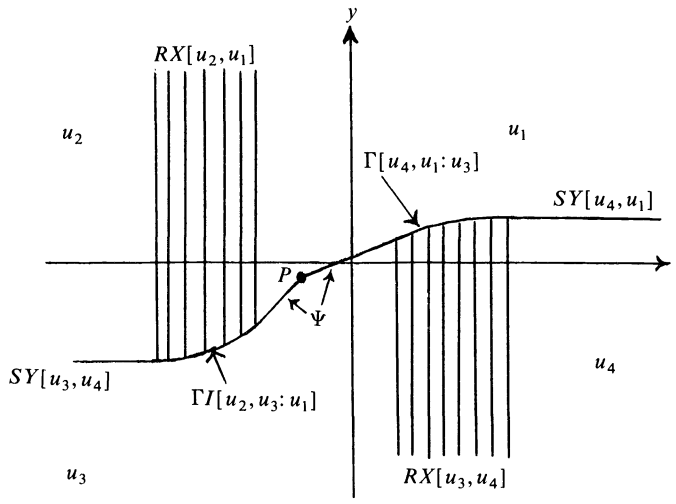


FIG. 5. *The case* $u_2<u_1<u_3<u_4$.

Note that the right edge of $RX[u_2,u_1]$ has equation $x=f'(u_1)t$. Since $f''>0$ and $u_1<u_3$, $f'(u_1)<f'(u_3)$; furthermore, $x=f'(u_3)t$ is the equation for the left edge of $RX[u_3,u_4]$. Thus $RX[u_2,u_1]$ lies completely to the left of $RX[u_3,u_4]$ in this case. The point P has the same coordinates as in Case 1.

Case 6. $u_4<u_1<u_3<u_2$. This is the reflection of Case 5. Note that Cases 5 and 6 are invariant under inversion.

Case 7. $u_2<u_3<u_1<u_4$. At $t=1$ the solution looks like Fig. 6. In this case, since $u_3<u_1$, and $f''>0$, we have that $f'(u_3)<f'(u_1)$; hence $RX[u_2,u_1]$ overlaps $RY[u_3,u_4]$. Also the shock waves $\Gamma[u_4,u_1]$ and $\Gamma I[u_2,u_3]$ are not truncated in this case. The point Q_1 has coordinates $(f'(u_1),g'(u_1))$, and $Q_2=(f'(u_3),g'(u_3))$.

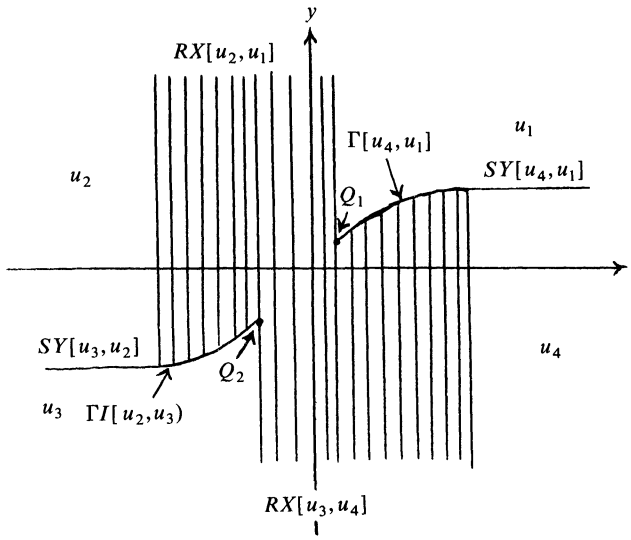


FIG. 6. *The case* $u_2<u_3<u_1<u_4$.

Case 8. $u_4<u_3<u_1<u_2$. This is the reflection of Case 7. Note that Cases 7 and 8 are invariant under inversion.

Case 9. $u_1<u_3<u_2<u_4$. At $t=1$ the solution looks like Fig. 7. The point P has the same coordinates as before.

Case 10. $u_1<u_3<u_4<u_2$. This is the reflection of Case 9.

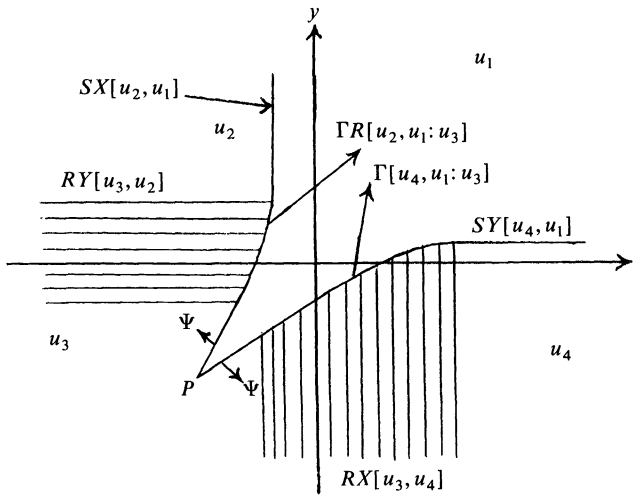


FIG. 7. *The case* $u_1<u_3<u_2<u_4$.

Case 11. $u_3>u_1>u_4>u_2$. This is the inversion of Case 9.
Case 12. $u_3>u_1>u_2>u_4$. This is the reflected inversion of Case 9.
Case 13. $u_1<u_2<u_4<u_3$. At $t=1$ the solution looks like Fig. 8.

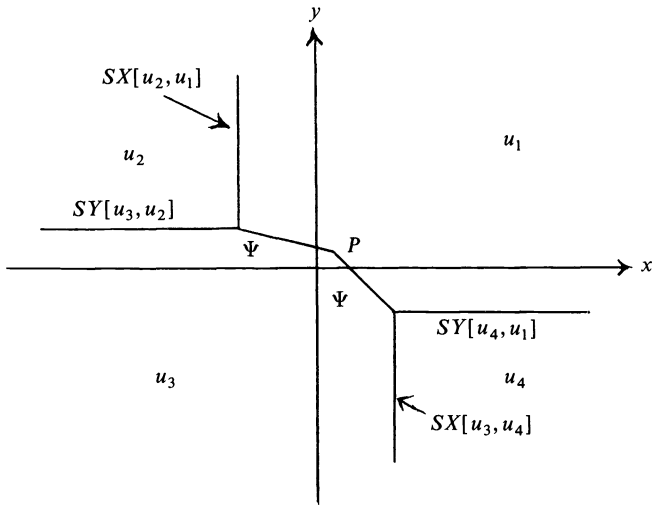


FIG. 8. *The case* $u_1<u_2<u_4<u_3$.

Case 14. $u_1<u_4<u_2<u_3$. This is the reflection, and also the inversion, of Case 13.

Case 15. $u_2<u_3<u_4<u_1$. At $t=1$ the solution looks like Fig. 9. Here Q has coordinates $(f'(u_3),g'(u_3))$. Between the points $C_1=(f'(u_4),g'(u_4))$ and $C_2=(f'(u_1),g'(u_1))$, the rarefaction waves $RX[u_4,u_1]$ and $RY[u_4,u_1]$ meet along the surface Δ, which may be described by the equations $x=f'(s)t$ and $y=g'(s)t$, for $u_4<s<u_1$ and $t>0$. The plane sections where $u=s$ in each wave meet along the line $x=f'(s)t, y=g'(s)t$. Thus the solution is continuous, though not differentiable, along this surface. One may check that the entropy condition is satisfied near Δ.

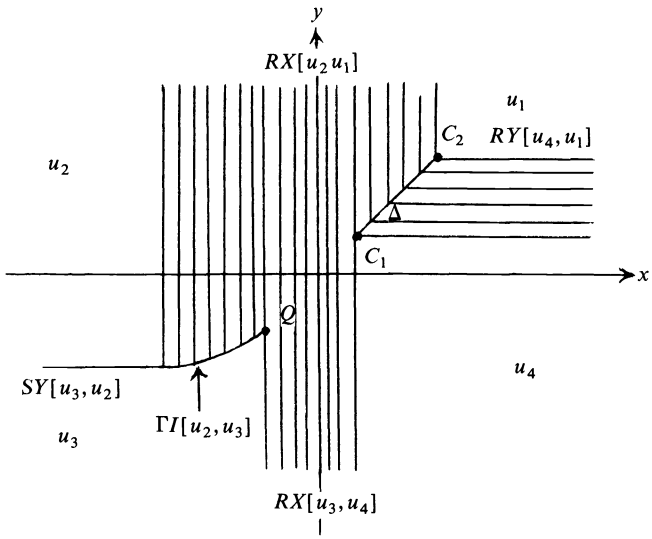


FIG. 9. *The case* $u_2<u_3<u_4<u_1$.

Case 16. $u_4<u_3<u_2<u_1$. This is the reflection of Case 15.
Case 17. $u_4>u_1>u_2>u_3$. This is the inversion of Case 15.
Case 18. $u_2>u_1>u_4>u_3$. This the reflected inversion of Case 15.
Case 19. $u_2<u_4<u_3<u_1$. At $t=1$ the solution looks like Fig. 10.

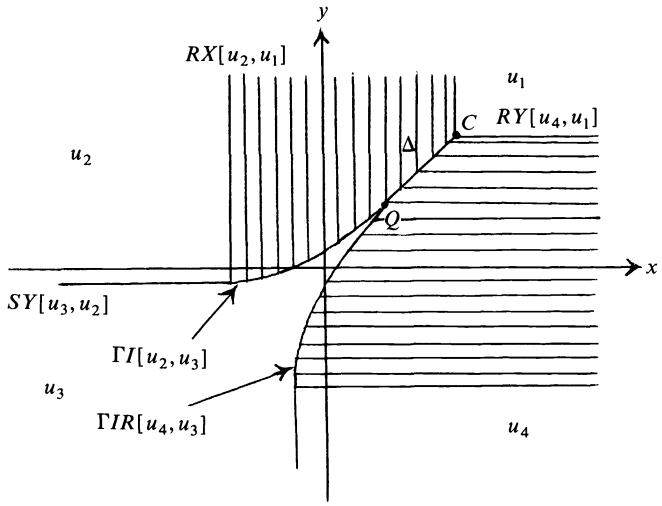


FIG. 10. *The case* $u_2<u_4<u_3<u_1$.

Case 20. $u_4<u_2<u_3<u_1$. This is the reflection of Case 19.
Case 21. $u_4>u_2>u_1>u_3$. This is the inversion of Case 19.
Case 22. $u_2>u_4>u_1>u_3$. This is the reflected inversion of Case 19.
Case 23. $u_3<u_2<u_4<u_1$. At $t=1$ the solution looks like Fig. 11. Here Q_1 has coordinates $(f'(u_3),g'(u_3))$. Also $Q_2=(f'(u_2),g'(u_2))$, $Q_3=(f'(u_4),g'(u_4))$, $Q_4=(f'(u_1),g'(u_1))$.

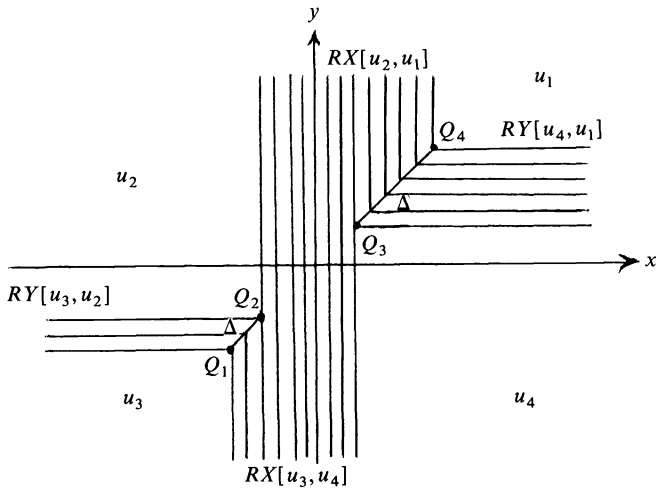


FIG. 11. *The case* $u_3<u_2<u_4<u_1$.

Case 24. $u_3<u_4<u_2<u_1$. This is the reflection of Case 23. Note that Case 23 and 24 are invariant under inversion. Also note that Cases 23 and 24 are the only cases with continuous solutions.

3. Verification of the entropy condition. We will now prove three theorems showing that under certain conditions, our solutions are single valued and satisfy the entropy condition. Theorem 1 treats the case where $f \equiv g$ and is convex. It is easily seen that this also includes the case $f'' = cg''$. Theorems 2 and 3 concern perturbations away from this case.

THEOREM 1. *If $f \equiv g$, then the constructions in* §2 *for Cases* 1–24 *define functions, and these functions satisfy the entropy condition.*

Remark. In Cases 9–12 and 19–22 it is conceivable that our solution may fail to be single valued due to overlapping Γ shocks, as illustrated in Fig. 12. Therefore it is necessary to prove that this does not occur.

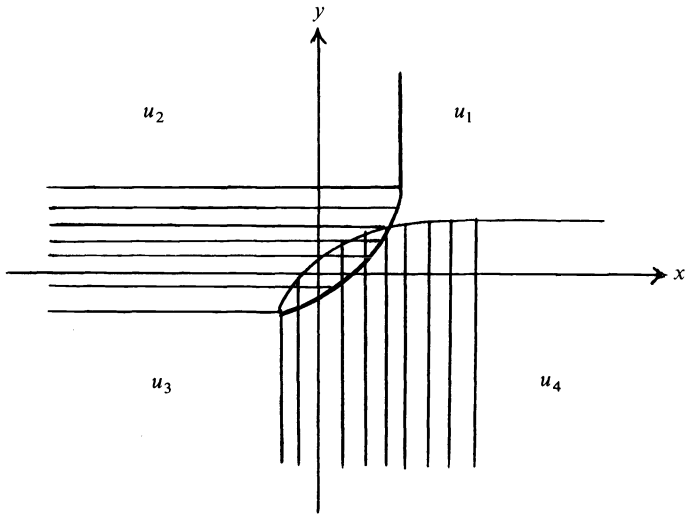


FIG. 12. *Overlap.*

Proof. We need to prove that the entropy condition is satisfied for SX, SY, Γ, and Ψ. However SX and SY are one dimensional shocks and the entropy condition is known to hold for them. The proof for $\Gamma[v,w]$ is as follows: To verify the entropy condition we require **n**, the normal vector to $\Gamma[v,w]$, to be oriented towards the side of the shock surface where u is larger. In this situation this means that **n** must "point in" to the rarefaction wave. Since the rarefaction wave lies in the region $y<\gamma(s)t$, we must have $n_y \leq 0$. Since $f''>0$, this means we should choose:

$$\mathbf{n} = (f''(s)\gamma(s) - f'(s)\gamma'(s), \gamma'(s), -f''(s)). \tag{3.1}$$

Thus, we must verify

$$(k-w)(f''(s)\gamma(s) - f'(s)\gamma'(s)) + (f(k)-f(w))(\gamma'(s)-f''(s)) \geq 0 \tag{3.2}$$

for all $w<k<s<v$.

LEMMA 1.1. *When $f \equiv g$, $\gamma'(s)<f''(s)$ for $s>w$.*

Proof. Using (2.13), and using the fact that $f''>0$, we have that $\gamma(v)<f'(v)$. Furthermore $\gamma = f'$ is a solution of (2.12). The uniqueness of solutions for $s>w$ implies that $\gamma(s)<f'(s)$ for all $s>w$. Using (2.12) and (2.13), one notes that $\gamma'(v)=0<f''(v)$, and also that $\gamma'(s)=f''(s)$ implies $f(w)-f(s)-\gamma(s)(w-s)=f(w)-f(s)-f'(s)(w-s)$; this in turn implies that $f'(s)=\gamma(s)$ when $s \neq w$. Since $\gamma(s)<f'(s)$, we must have that $\gamma'(s)<f''(s)$ for $s>w$.

We now verify (3.2). If the left-hand side of (3.2) is considered as a function $F(k)$, then one checks using Lemma 1.1 that $F''(k)>0$; also $F(w)=0$. Furthermore, $F(s)=0$ since in (3.2) one may replace all w's by s's, using (2.1). Therefore $F(k)>0$ for $w<k<s<v$; this is (3.2).

The entropy condition for reflected and inverted $\Gamma[v,w]$ shock waves may similarly be verified.

It remains to verify the entropy condition for Ψ-shocks. These appear in two different contexts:

(1) Tangential shocks, that is, Ψ-shocks which are tangential to Γ-shocks. Such a Ψ shock has the same normal vector $\mathbf{n}$ as does the Γ-shock at the point of tangency, and also the same values of u on either side of the shock. Therefore the entropy condition for a tangential Ψ-shock is equivalent to the entropy condition for the corresponding Γ shock at the point of tangency.

(2) Nontangential shocks. One may check, using the ordering of u_1, u_2, u_3, and u_4, and the convexity of f and g, that the intersection of any nontangential Ψ-shock surface with the plane $t=1$ is a line segment with negative slope, as depicted in Figs. 4 and 8. Furthermore any nontangential Ψ-shock is a shock between u_3 and u_1, with $u_3>u_1$, with the region $u=u_1$ above, and to the right of the shock. Thus $\mathbf{n}=(n_t,n_x,n_y)$ with n_x and n_y negative. The entropy condition becomes

$$(k-u_3)n_t+(f(k)-f(u_3))(n_x+n_y)\geq 0, \tag{3.3}$$

for $u_3>k>u_1$. Since $n_x+n_y<0$, and f is convex, the left side of this inequality has negative second derivative with respect to k, and by (2.1), is zero when $k=u_3$ or $k=u_1$. Hence it is positive for values of k between u_1 and u_3.

To verify that our solutions are single valued it suffices to show that overlap of Γ-shocks does not occur. In the proof of Lemma 1.1 we saw that $\gamma(s)<f'(s)$; thus both Γ shocks lie on opposite sides of the curve $x=f'(s)$, $y=f'(s)$.

THEOREM 2. *Let u_1, u_2, u_3, u_4 be such that our proposed solution to* (1.1), (1.2) *contains no complete Γ, ΓR, ΓI, or ΓIR shock, that is, let us consider only Cases* 1–6, 9–14, 23, *and* 24 *of* §2. *Let $M=\max_{1\leq i\leq 4}u_i$, and $m=\min_{1\leq i\leq 4}u_i$. Then for any given function h such that $h''>0$, there exists $\varepsilon>0$ such that whenever $\|f-h\|_{C^2[m,M]}<\varepsilon$, and $\|g-h\|_{C^2[m,M]}<\varepsilon$, our construction for the solution to* (1.1), (1.2) *in these cases defines a function, and this function satisfies the entropy condition.*

Proof. As noted before, it suffices to prove that in our solution to the perturbed equation, the entropy condition holds for truncated Γ-shocks, and for nontangential Ψ-shocks.

Recall that the entropy condition states that for $\Gamma[v,w:p]$ we must have:

$$(k-w)n_t+(f(k)-f(w))n_x+(g(k)-g(w))n_y\geq 0, \tag{3.4}$$

or

$$\begin{aligned} W(k,s)=&(f''(s)\gamma(s)-f'(s)\gamma'(s))(k-w)\\ &+\gamma'(s)(f(k)-f(w))-f''(s)(g(k)-g(w))\geq 0, \end{aligned} \tag{3.5}$$

for all $w\leq k\leq s\leq v$, and $w<p\leq s$. The function γ satisfies the initial value problem (2.12), (2.13). Since f is convex (for ε sufficiently small), and $s\geq p>w$, the denominators in (2.14), the formula for γ, are bounded away from zero, and the bound depends only on h,ε,v,w and p. Thus we may conclude that the map $(f,g)\to\gamma$ is continuous from $\{f|f\in C^2[p,v],\ f''>0\}x\{g|g\in C^2[p,v],\ g''>0\}$ to $C^1[p,v]$.

Next note that

$$\frac{\partial^2 W}{\partial k^2}=\gamma'(s)f''(k)-f''(s)g''(k), \tag{3.6}$$

and note that $W(s,s)=W(w,s)=0$, by (2.1). When $f\equiv g\equiv h$ we know by Lemma 1.1 that $\gamma'(s)<f''(s)$ for $p<s<v$. Thus $\partial^2 W/\partial k^2<0$ for $w\le k\le s\le v$, $w<p<s$, when $f\equiv g\equiv h$. Moreover, $\partial^2 W/\partial k^2$ depends continuously on f and g in the C^2 topology on f and g. Therefore there exists $\varepsilon>0$ such that $\|f-h\|_{C^2}<\varepsilon$ and $\|g-h\|_{C^2}<\varepsilon$ imply $\partial^2 W/\partial k^2>0$, and it follows that $\Gamma[v,w:p]$ satisfies the entropy condition.

Since $\gamma(s)<f'(s)$ when $f\equiv g\equiv h$, for ε sufficiently small we must have $\gamma(s)<g'(s)$; thus in Case 9–12 the two Γ-shocks lie on opposite sides of the surface $x=f'(s)t, y=g'(s)t$. Thus overlap does not occur for ε sufficiently small.

Finally, for nontangential Ψ-shocks we must show:

$$(k-u_3)n_t+(f(k)-f(u_3))n_x+(g(k)-g(u_3))n_y\ge 0 \tag{3.7}$$

for $u_1\le k\le u_3$. Since $n_x<0$ and $n_y<0$ as we saw in the previous section, and since $f''>0$ and $g''>0$ for ε sufficiently small, the left-hand side of (3.7) has negative second derivative with respect to k, and equals zero when $k=u_1$ or $k=u_3$. Thus (3.7) holds for values of k between u_1 and u_3.

THEOREM 3. *Suppose u_1, u_2, u_3, and u_4 are such that our proposed formula for the solution to* (1.1), (1.2), *contains some complete* $\Gamma[v,w]$, $\Gamma R[v,w]$, $\Gamma I[v,w]$, *or* $\Gamma IR[v,w]$ *shock, that is, let us consider Cases* 7, 8 *and* 15–22; *suppose also that $f''(w)=g''(w)$ and $f'''(w)=g'''(w)$ at all w which occur as above. (Note: w is always either u_1 or u_3.) Let m and M be as defined in Theorem* 2. *Then for any given function h such that $h''>0$ and $h''(w)=f''(w)$, $h'''(w)=f'''(w)$, there exists $\varepsilon>0$ such that if $\|f-h\|_{C^4[m,M]}<\varepsilon$, $\|g-h\|_{C^4[m,M]}<\varepsilon$, then our proposed solution is single valued, and it satisfies the entropy condition.*

Proof. The proof consists of several lemmas.

LEMMA 3.1. *The map T: $C^4[v,w]\times C^4[v,w]\to C^1[v,w]$, $T(f,g)=\gamma$, where γ satisfies* (2.12), (2.13), *is continuous for $f''>0$, $g''>0$. The map $S_s(f,g)=(d/ds)T(f,g)(s)=\gamma'(s)$ satisfies*

$$|(DS_s)_{(f,g)}(p,q)|\le C|s-w|\|(p,q)\|_{C^4[v,w]}, \tag{3.8}$$

for all p,q in $C^4[v,w]$ such that $p''(w)=q''(w)=p'''(w)=q'''(w)=0$, where $C>0$ does not depend on p,q, or s.

Proof. From (2.14) we have

$$\begin{aligned}\gamma(s)&=\frac{g(s)-g(w)}{s-w}\\&\quad-\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)\,dz}{f(w)-f(z)-f'(z)(w-z)}\right)\frac{g(w)-g(r)-g'(r)(w-r)}{(w-r)^2}\,dr\\&=(def.)\ T_s(f,g).\end{aligned}$$

Clearly T_s is a continuous and differentiable mapping from $\{f\,|\,f\in C^4[w,w],\ f''>0\}\times\{g\,|\,g\in C^4[v,w],\ g''>0\}$ to $\mathbb{R}$, for $s\in(w,v]$. However, it is necessary to show that $DT_{s(f,g)}$ is bounded independent of s, and independent of the choice of (f,g) from a neighborhood of (h,h).

Define $T_w(f,g)$ to equal $g'(w)$. We prove that $s\to T_s(f,g)$ is continuous at w, as follows: We have

$$\gamma(s)=T_s(f,g)=\frac{g(s)-g(w)}{s-w}-\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)\,dz}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\right)\frac{\int_w^r g''(\theta)(\theta-w)\,d\theta}{(w-r)^2}\,dr.$$

Note that

$$\exp\left(\int_s^r \frac{f''(z)(w-z)\,dz}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\right)$$

$$=\exp\left(\int_s^r \frac{2}{w-z}+\frac{2\int_w^z f'''(\theta)(\theta-w)^2 d\theta}{(w-z)\left(f''(z)(z-w)^2-\int_w^z f'''(\theta)(\theta-w)^2 d\theta\right)}\,dz\right)$$

$$=\frac{(s-w)^2}{(r-w)^2}\exp\left(\int_s^r \frac{\int_w^z f'''(\theta)(\theta-w)^2 d\theta}{(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right).$$

Thus

$$\left|\gamma(s)-\frac{g(s)-g(w)}{s-w}\right|$$

$$<(s-w)^2\int_s^v \exp\left(\int_s^r \frac{2\|f'''\|_\infty}{3(\inf(f''))}\,dz\right)\frac{\|g''\|_\infty}{(r-w)^2}\,dr$$

$$=(s-w)^2\left(\frac{1}{s-w}-\frac{1}{v-w}\right)\frac{\|g''\|_\infty}{2}\exp\left(\frac{2\|f'''\|_\infty(v-w)}{3(\inf(f''))}\right),$$

and so we see that $\lim_{s\to w}\gamma(s)=g'(w)$; thus $s\to T_s(f,g)$ is continuous on $[v,w]$. The sup norms and infimums used above and henceforth are all taken over $[v,w]$.

Next,

$$(DT_s)_{(f,g)}(p,q)$$

$$=\frac{q(s)-q(w)}{s-w}-\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right)$$

$$\cdot\left[\left(\int_s^r \frac{p''(z)(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta-f''(z)(z-w)\int_w^z p''(\theta)(\theta-w)\,d\theta}{\left[\int_w^z f''(\theta)(\theta-w)\,d\theta\right]^2}\,dz\right)\right.$$

$$\left.\cdot\int_w^r \frac{g''(\theta)(\theta-w)}{(r-w)^2}\,d\theta+\int_w^r \frac{q''(\theta)(\theta-w)}{(r-w)^2}\,d\theta\right]dr.$$

Thus

$$|(DT_s)_{(f,g)}(p,q)|$$

$$\le\|q'\|_\infty+\int_s^v \frac{(s-w)^2}{(r-w)^2}\exp\left(\frac{2\|f'''\|_\infty(v-w)}{3(\inf(f''))}\right)$$

$$\cdot\left|\left[\left(\int_s^r\left[\int_w^z f''(\theta)(\theta-w)\,d\theta\right]^{-2}\right.\right.\right.$$

$$\cdot\left[\left(p''(w)+\int_w^z p'''(\theta)\,d\theta\right)(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta\right.$$

$$-\frac{1}{2}f''(z)(z-w)\left(p''(w)(z-w)^2\right.$$

$$\left.\left.-\int_w^z p'''(\theta)\left((\theta-w)^2-(z-w)^2\right)d\theta\right)\right]dz\Bigg)\frac{\|g''\|_\infty}{2}$$

$$\left.+\frac{1}{2}\frac{q''(w)(r-w)^2-\int_w^r q'''(\theta)\left[(\theta-w)^2-(r-w)^2\right]d\theta}{(r-w)^2}\right]\Bigg|dr.$$

Since only perturbations fixing $f''(w)$, $g''(w)$, $f'''(w)$, and $g'''(w)$ are considered, we have $p''(w)=q''(w)=p'''(w)=q'''(w)=0$. Thus

$$\begin{aligned}
&\left|(DT_s)_{(f,g)}(p,q)\right| \\
&\quad\le\|q'\|_\infty+\int_s^v\frac{(s-w)^2}{(r-w)^2}\exp\left(\frac{2\|f'''\|_\infty(v-w)}{3(\inf(f''))}\right) \\
&\qquad\cdot\left[\frac{\|p'''\|_\infty\|f''\|_\infty}{(\inf(f''))^2}\left(\frac{\frac{1}{2}+\frac{2}{3}}{\frac{1}{4}}\right)\frac{\|g''\|_\infty}{2}+\|q'''\|_\infty\frac{(r-w)}{3}\right]dr \\
&\quad\le\|q'\|_\infty+\exp\left(\frac{2\|f'''\|_\infty(v-w)}{3(\inf(f''))}\right) \\
&\qquad\cdot\left[\frac{\|p'''\|_\infty\|f''\|_\infty\|f''\|_\infty\|g''\|_\infty}{(\inf(f''))^2}\left(\frac{5}{3}\right)\left(\frac{1}{s-w}-\frac{1}{v-w}\right)\right. \\
&\qquad\left.+\frac{1}{3}\|q'''\|_\infty(\ln(v-w)-\ln(s-w))\right]\cdot(s-w)^2,
\end{aligned}$$

and so we see that $\|(DT_s)_{(f,g)}\|$ is independent of s, $w\le s\le v$, and also independent of the choice of (f,g) from a sufficiently small neighborhood of (h,h). Therefore by the mean value theorem the map $(f,g,s)\to\gamma(s)$ is locally Lipschitz for each s, with a Lipschitz constant independent of s. Thus $(f,g)\to\gamma$ is a continuous mapping from $C^3[w,v]\times C^3[w,v]\to C[w,v]$. Next,

$$\begin{aligned}
&S_s(f,g)=\gamma'(s) \\
&\quad=\frac{g'(s)(s-w)-(g(s)-g(w))}{(s-w)^2}-\frac{g(w)-g(s)-g'(s)(w-s)}{(s-w)^2} \\
&\qquad-\int_v^s\exp\left(\int_s^r\frac{f''(z)(w-z)}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\right)\left(\frac{-f''(s)(w-s)}{\int_w^s f''(\theta)(\theta-w)\,d\theta}\right)\frac{\int_w^r g''(\theta)(\theta-w)\,d\theta}{(r-w)^2}dr.
\end{aligned}$$

Differentiating with respect to f and g, we have

$$
\begin{aligned}
&(DS_s)_{(f,g)}(p,q)\\
&=(w-s)\,\frac{\left[p''(s)\int_w^s f''(\theta)(\theta-w)\,d\theta-f''(s)\int_w^s p''(\theta)(\theta-w)\,d\theta\right]}{\left[\int_w^s f''(\theta)(\theta-w)\,d\theta\right]^2}\\
&\qquad\cdot\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right)\frac{\int_w^r g''(\theta)(\theta-w)\,d\theta}{(r-w)^2}\,dr\\
&\quad+\frac{f''(s)(w-s)}{\int_w^s f''(\theta)(\theta-w)\,d\theta}\\
&\qquad\cdot\int_v^s \exp\left(\int_s^r \frac{f''(z)(w-z)}{\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right)\\
&\qquad\cdot\left[\left(\int_s^r\left[\int_w^z f''(\theta)(\theta-w)\,d\theta\right]^{-2}\right.\right.\\
&\qquad\qquad(w-z)\left[p''(z)\int_w^z f''(\theta)(\theta-w)\,d\theta-f''(z)\int_w^z p''(\theta)(\theta-w)\,d\theta\right]dz\Bigg)\\
&\qquad\qquad\cdot\int_w^r\frac{g''(\theta)(\theta-w)}{(r-w)^2}\,d\theta+\int_w^r\frac{q''(\theta)(\theta-w)}{(r-w)^2}\,d\theta\Bigg]\,dr\\
&=\frac{(w-s)\left[\int_w^s p'''(\theta)\,d\theta\int_w^s f''(\theta)(\theta-w)\,d\theta+\frac{f''(s)}{2}\int_w^s p'''(\theta)\left((\theta-w)^2-(s-w)^2\right)d\theta\right]}{\left[\int_w^s f''(\theta)(\theta-w)\,d\theta\right]^2}\\
&\qquad\cdot\int_v^s\left(\frac{s-w}{r-w}\right)^2\exp\left(\int_s^r\frac{\int_w^z f'''(\theta)(\theta-w)^2d\theta}{(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right)\int_w^r\frac{g''(\theta)(\theta-w)\,d\theta}{(r-w)^2}\,dr\\
&\quad+\frac{f''(s)(w-s)}{\int_w^s f''(\theta)(\theta-w)\,d\theta}\\
&\qquad\cdot\int_v^s\exp\left(\int_s^r\frac{\int_w^z f'''(\theta)(\theta-w)^2d\theta}{(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta}\,dz\right)\left(\frac{s-w}{r-w}\right)^2\\
&\qquad\cdot\left[\int_s^r(w-z)\left[\int_w^z f''(\theta)(\theta-w)\,d\theta\right]^{-2}\right.\\
&\qquad\qquad\cdot\left[\int_w^z p'''(\theta)\,d\theta\int_w^z f''(\theta)(\theta-w)\,d\theta\right.\\
&\qquad\qquad\qquad+f''(z)\int_w^z\frac{p'''(\theta)}{2}\left((\theta-w)^2-(z-w)^2\right)d\theta\Bigg]\,dz\\
&\qquad\qquad\cdot\int_w^r\frac{g''(\theta)(\theta-w)}{(r-w)^2}\,d\theta-\int_w^r\frac{q'''(\theta)\left((\theta-w)^2-(r-w)^2\right)}{2(r-w)^2}\,d\theta\Bigg]\,dr.
\end{aligned}
$$

Note that, since $p'''(w)=0$,

$$\int_w^s p'''(\theta)\,d\theta=-\int_w^s p''''(\theta)((\theta-w)-(s-w))\,d\theta,$$

and

$$\int_w^z p'''(\theta)(\theta-w)^2 d\theta=\int_w^z \frac{1}{3}p''''(\theta)\big((z-w)^3-(\theta-w)^3\big)\,d\theta.$$

Thus

$$\begin{aligned}
&(DS_s)_{(f,g)}(p,q)\\
&\quad=(w-s)\left[\int_w^s f''(\theta)(\theta-w)\,d\theta\right]^{-2}\\
&\qquad\cdot\left[\int_w^s p''''(\theta)((s-w)-(\theta-w))\,d\theta\int_w^s f''(\theta)(\theta-w)\,d\theta\right.\\
&\qquad\left.+f''(s)\int_w^s p''''(\theta)\left[\frac{(s-w)^3}{6}-\frac{(\theta-w)^3}{6}-\frac{(s-w)^2}{2}((s-w)-(\theta-w))\right]d\theta\right]\\
&\qquad\cdot\int_v^s\left(\frac{s-w}{r-w}\right)^2\exp\left(\int_s^r\frac{\int_w^z f'''(\theta)(\theta-w)^2d\theta}{(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta}dz\right)\cdot\int_w^r\frac{g''(\theta)(\theta-w)}{(r-w)^2}d\theta\,dr\\
&\qquad+\frac{f''(s)(w-s)}{\int_w^s f''(\theta)(\theta-w)\,d\theta}\\
&\qquad\cdot\int_v^s\left(\frac{s-w}{r-w}\right)^2\exp\left(\int_s^r\frac{\int_w^z f'''(\theta)(\theta-w)^2d\theta}{(w-z)\int_w^z f''(\theta)(\theta-w)\,d\theta}dz\right)\\
&\qquad\cdot\left[\left(\int_s^r(w-z)\left(\int_w^z f''(\theta)(\theta-w)\,d\theta\right)^{-2}\right.\right.\\
&\qquad\cdot\left(\int_w^z p''''(\theta)((z-w)-(\theta-w))\,d\theta\int_w^z f''(\theta)(\theta-w)\,d\theta\right.\\
&\qquad+f''(z)\int_w^z p''''(\theta)\left[\frac{(z-w)^3}{6}-\frac{(\theta-w)^3}{6}\right.\\
&\qquad\left.\left.\left.-\frac{(z-w)^2}{2}((z-w)-(\theta-w))\right]d\theta\right)dz\right)\\
&\qquad\cdot\int_w^r\frac{g''(\theta)(\theta-w)}{(r-w)^2}d\theta\\
&\qquad+\int_w^r q''''(\theta)\left[\frac{(r-w)^3}{6}-\frac{(\theta-w)^3}{6}-\frac{(r-w)^2}{2}((r-w)-(\theta-w))\right]d\theta\Bigg]dr.
\end{aligned}$$

So

$$
\begin{aligned}
&|(DS_s)_{(f,g)}(p,q)| \\
&\quad \le \frac{\|p''''\|_\infty \|f''\|_\infty}{(\inf(f''))^2}(5)|s-w|^3 \\
&\qquad \cdot \left|\frac{1}{v-w}-\frac{1}{s-w}\right| \exp\left(\frac{2\|f'''\|_\infty (v-w)}{3(\inf(f''))}\right)\cdot \frac{\|g''\|_\infty}{2} \\
&\qquad + \frac{\|f''\|_\infty}{(\inf(f''))^2}|s-w| \exp\left(\frac{2\|f'''\|_\infty (v-w)}{3(\inf(f''))}\right) \\
&\qquad\quad \cdot \left[\frac{\|p''''\|_\infty \|f''\|_\infty \|g''\|_\infty}{2(\inf(f''))^2}\cdot(5)\int_s^v \frac{1}{2(r-w)^2}\left|(r-w)^2-(s-w)^2\right| dw \right. \\
&\qquad\qquad \left. + \int_s^v \frac{1}{(r-w)^2}\|q''''\|_\infty \cdot \frac{1}{8}\cdot (s-w)^2 dw\right].
\end{aligned}
$$

Thus $\|(DS_s)(f,g)\| \le C|s-w|$ for some $C>0$, where C depends only on $\|f'''\|_\infty$, $\inf(f'')$, $\|g''\|_\infty$, and (v,w). □

LEMMA 3.2. *In a $\Gamma[v,w]$ shock, in the case where $f\equiv g$, $\gamma''(w)<f'''(w)$.*

Proof. Recall that when $f\equiv g$, f' is a solution to (2.12). Thus $\gamma(s)=f'(s)+\sigma(s)$, where σ satisfies

$$\sigma'(s)=\sigma(s)\frac{f''(s)(s-w)}{\int_w^s f''(\theta)(\theta-w)\,d\theta}.$$

Thus

$$(s-w)\frac{d}{ds}\ln(\sigma(s))=\frac{f''(s)(s-w)^2}{\int_w^s f''(\theta)(\theta-w)\,d\theta}.$$

Note that

$$\lim_{s\to w}\frac{f''(s)(s-w)^2}{\int_w^s f''(\theta)(\theta-w)\,d\theta}=2.$$

Thus one may write

$$\sigma(s)=C(s-w)^2\exp\left(\int_v^s \frac{\int_w^z f'''(\theta)(\theta-w)^2 d\theta}{(z-w)\int_w^z f''(\theta)(\theta-w)\,d\theta}dz\right).$$

One may observe that σ has 2 continuous derivatives at w if f has 3 continuous derivatives, and $\sigma''(w)<0$ if and only if $C<0$.

Thus $\gamma''(w)<f'''(w)$ if and only if $\gamma<f'$. However, $\gamma(v)=(f(v)-f(w))/(v-w)<f'(v)$ since $f''<0$.

We can now prove that for sufficiently small ε, $\|f-h\|_{C^4}<\varepsilon$ and $\|g-h\|_{C^4}<\varepsilon$, together with the hypotheses of Theorem 3, imply that overlap does not occur. It suffices to show that for ε sufficiently small, $\gamma(s)<f'(s)$, $w<s\le v$. From this it follows that the line $y=x$ separates the two Γ-shocks in Cases 19 through 22.

Since when $f\equiv g\equiv h$, $\gamma(s)<f'(s)$ for $w<s\leq v$, and since γ depends continuously on f and g, we have that for any $\delta>0$ there exists $\varepsilon>0$ such that $\|f-h\|_{C^4}<\varepsilon$ and $\|g-h\|_{C^4}<\varepsilon$ imply $\gamma(s)<f'(s)$ for $w+\delta\leq s\leq v$. For s near w, and $f\equiv g\equiv h$,

$$f''(s)-\gamma'(s)=(f'''(w)-\gamma''(w))(s-w)+o(s-w).$$

Since by Lemma 3.2 $f'''(w)-\gamma''(w)>0$, for δ sufficiently small we have

$$f''(s)-\gamma'(s)>C_0(s-w)$$

for some $C_0>0$ and for $w\leq s\leq\delta$. By Lemma 3.1 and the mean value theorem, for $\|f-h\|_{C^4}<\varepsilon$ and $\|g-h\|_{C^4}<\varepsilon$

$$f''(s)-\gamma'(s)>(C_0-\varepsilon C)(s-w)$$

for $w\leq s<\delta$. For ε sufficiently small $C_0-\varepsilon C>0$, and hence $f''(s)>\gamma'(s)$ for $w<s\leq\delta$. Since $\gamma(w)=f'(w)$, the result follows.

To finish the proof of Theorem 3, it must be shown that

$$W(k,s)=(f''(s)\gamma(s)-f'(s)\gamma(s))(k-w)$$
$$+\gamma'(s)(f(k)-f(w))-f''(s)(g(k)-g(w))\geq 0,$$

for all $k,s,w\leq k\leq s\leq v$. Recall that $W(s,s)=W(w,s)=0$, and, when $f\equiv g$, $\partial^2W/\partial k^2=f''(k)(\gamma'(s)-f''(s))$ which is less than zero for $w<s\leq v$. In fact, for s close to w,

$$\left(\frac{\partial^2W}{\partial k^2}\right)=f''(w)(\gamma''(w)-f'''(w))(s-w)+o(s-w).$$

On the other hand,

$$D_{(f,g)}\left(\frac{\partial^2W}{\partial k^2}\right)(p,q)$$
$$=D_{(f,g)}(\gamma')(p,q)f''(k)+\gamma'(s)p''(k)-p''(s)g''(k)-f''(s)q''(k).$$

Thus

$$\left|D_{(f,g)}\left(\frac{\partial^2W}{\partial k^2}\right)(p,q)\right|$$
$$\leq O(s-w)\|(p,q)\|_{C^4}f''(k)+|\gamma'(s)p''(k)-p''(s)g''(k)-g''(s)q''(k)|.$$

Since we are considering only those tangent vectors p and q such that $p''(w)=q''(w)=0$, we have that $|p''(k)|$, $|p''(s)|$, and $|q''(k)|$ are less than $(\|p\|_{C^4}+\|q\|_{C^4})|s-w|$. Thus for some $C_1>0$,

$$\left|D_{(f,g)}\left(\frac{\partial^2W}{\partial k^2}\right)(p,q)\right|\leq C_1\cdot|s-w|\cdot\|(p,q)\|_{C^4}.$$

Thus for $f\equiv g\equiv h$, $\partial^2W/\partial k^2\leq C_0\cdot|s-w|$ for some $C_1>0$; and for $f=h+\varepsilon p$, $g=h+\varepsilon q$, where p and q satisfy $p''(w)=q''(w)=p'''(w)=q'''(w)=0$,

$$\frac{\partial^2W}{\partial k^2}\leq\big(\varepsilon(C+C_1)\|(p,q)\|_{C^4}-C_0\big)|s-w|\leq 0$$

for sufficiently small ε. Since under these conditions W is convex down, and $W=0$ at $k=s$ and $k=w$, one concludes that $W\geq 0$ for $w\leq k\leq s\leq v$. □

Remark. In Theorems 2 and 3 one may, of course, fix f and perturb only g, or the other way around. In this case one chooses $h\equiv f$, or $h\equiv g$, respectively.

4. A counterexample. The following is an example of a C^∞ one-parameter family (f,f_ε) of pairs of C^∞ functions such that $f=f_0$ and such that for certain initial data, the solution given in §2 does not satisfy the entropy condition for $\varepsilon>0$. However, we also give the correct entropy solution for this example.

Let $f(s)=s^2$, and $f_\varepsilon(s)=s^2+\varepsilon s^3$. Consider the initial value problem

$$\begin{aligned} &\frac{\partial}{\partial t}u(t,x,y)+\frac{\partial}{\partial x}f(u(t,x,y))+\frac{\partial}{\partial y}f_\varepsilon(u(t,x,y))=0,\\ &u(0,x,y)=\begin{cases}\delta & \text{for } x>0 \text{ and } y<0,\\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{4.1}$$

where $\delta>0$. Note that $f_\varepsilon''>0$ on $[0,\delta]$ for $\varepsilon>-1/(3\delta)$. The solution given in §2 is described in Fig. 13.

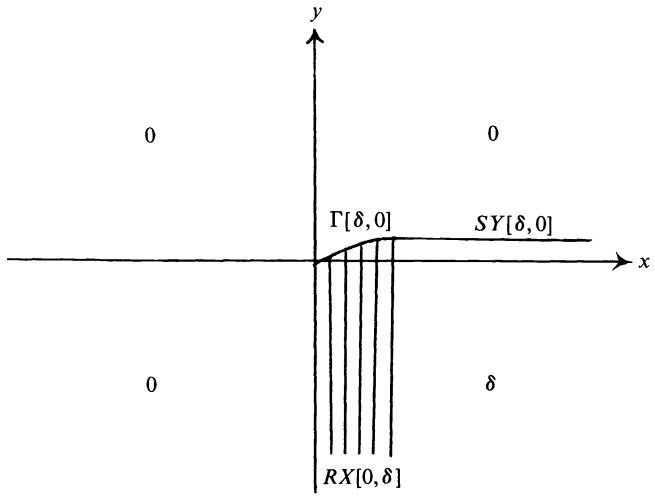


FIG. 13.

PROPOSITION. *If* $f(s)=s^2$, $g(s)=f_\varepsilon(s)=s^2+\varepsilon s^3$, *then for no choice of* $\varepsilon>0$, *and* δ *such that* $0<\delta$ *does the shock* $\Gamma[\delta,0]$ *satisfy the entropy condition, near the line* $x=0$, $y=0$.

Proof. $\Gamma[\delta,0]$ is described by the equations $x=f'(s)t$, $y=\gamma(s)t$, $0\le s\le\delta$, where γ satisfies

$$\begin{aligned} &\gamma'(s)=2\,\frac{0-s^2-\varepsilon s^3-\gamma(s)(0-s)}{0-s^2-2s(0-s)}=2\,\frac{\gamma(s)s-s^2-\varepsilon s^3}{s^2},\\ &\gamma(\delta)=\frac{\delta^2+\varepsilon\delta^3}{\delta}=\delta+\varepsilon\delta^2.\end{aligned} \tag{4.2}$$

Thus one may solve to find

$$\gamma(s)=2s-2\varepsilon s^2\ln(s)+s^2\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right). \tag{4.3}$$

Also

$$\gamma'(s)=2-4\varepsilon s\ln(s)-2\varepsilon s+2s\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right). \tag{4.4}$$

The shock wave $\Gamma[\delta,0]$ satisfies the entropy condition if

$$W(k,s)=(k-0)(2\gamma(s)-2s\gamma'(s))+(k^2-0)\gamma'(s)-(k^2+\varepsilon k^3)\cdot 2 \geq 0 \quad \text{for } 0\leq k\leq s\leq \delta. \tag{4.5}$$

Note that $\frac{\partial W}{\partial k}(0,s)=2\gamma(s)-2s\gamma'(s)$. Solving $\gamma(s)-s\gamma'(s)=0$, we have

$$\begin{aligned} &2s-2\varepsilon s^2\ln(s)+s^2\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right)\\ &\quad-\left[2s-4\varepsilon s^2\ln(s)-2\varepsilon s^2+2s^2\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right)\right]=0,\\ &\quad 2\varepsilon s^2\ln(s)+2\varepsilon s^2-s^2\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right)=0,\\ &\quad s^2\left[2\varepsilon\ln(s)+2\varepsilon-\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right)\right]=0. \end{aligned} \tag{4.6}$$

Thus $\gamma(s)=s\gamma'(s)$ if $s=0$ or

$$\begin{aligned} s&=\exp\left(-\frac{1}{2\varepsilon}\left(2\varepsilon-\left(-\frac{1}{\delta}+\varepsilon(2\ln(\delta)+1)\right)\right)\right)\\ &=(\text{def.})u_0. \end{aligned} \tag{4.7}$$

Note that u_0 is positive, and that $\lim_{\varepsilon\to 0}u_0=0$ for $0<\delta$. Furthermore, for $0<s<u_0$, $\gamma(s)-s\gamma'(s)<0$, hence, $\frac{\partial W}{\partial k}(0,s)<0$ for these values of s, and thus $W<0$ for these values of s, and k close to zero. Thus for any choice of δ greater than 0 our solution does not satisfy the entropy condition for small ε.

To keep the computations relatively simple, we give the correct entropy solution for this example only in the case $\delta=1$. In this case we will see that $\Gamma[1,0:u_0]$ satisfies the entropy condition. To the left of $x=f'(u_0)t$, a new rarefaction wave appears, and the shock wave passes between this new rarefaction wave and $RX[0,1]$. At $t=1$ the solution looks like Fig. 14.

Let us call the new rarefaction wave Σ, and the continuation of $\Gamma[1,0:u_0]$, Ω. Let Ω have equations $x=f'(s)t$, $y=w(s)t$. Below Ω, the solution u equals s on plane sections $x=f'(s)t$. Above Ω, $u=v$ on plane sections

$$y-f_\varepsilon'(v)t=\frac{w'(s)}{f''(s)}(x-f'(v)t). \tag{4.8}$$

These plane sections meet the shock curve at $x=f'(s)t$, $y=w(s)t$ tangentially, as suggested by (4.8).

We now prove that this description is correct, and give an explicit expression for w and the relationship between s and v.

Since Ω is to satisfy Condition 2.1, we have

$$w'(s)=f''(s)\,\frac{f_\varepsilon(v)-f_\varepsilon(s)-w(s)(v-s)}{f(v)-f(s)-f'(s)(v-s)}. \tag{4.9}$$

Furthermore, by hypothesis, $u=v$ along a plane tangent to $x=f'(s)t$, $y=w(s)t$. Thus the following result, due to Guckenheimer [5], will allow us to get a different expression for $w'(s)$.

THEOREM. *If u is a solution to the Riemann problem, then inside each region of rarefaction, the surfaces $u=v$ are sections of planes passing through the line $x=f'(v)t$, $y=f_\varepsilon'(v)t$.*

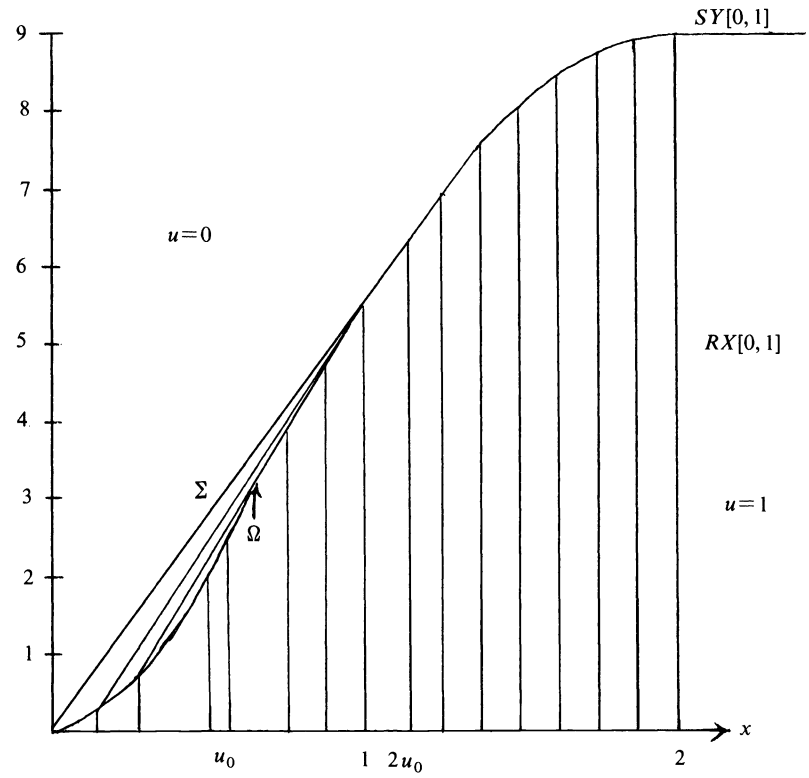


FIG. 14. *The solution to* (4.1) *with* $\delta=1, \varepsilon=8$.

From this result, we get

$$\frac{dy}{dx}=\frac{w'(s)}{f''(s)}=\frac{\Delta y}{\Delta x}=\frac{w(s)-f_\varepsilon'(v)}{f'(s)-f'(v)}. \tag{4.10}$$

Thus

$$\frac{v^2+\varepsilon v^3-s^2-\varepsilon s^3-w(s)(v-s)}{v^2-s^2-2s(v-s)}=\frac{w(s)-2v-3\varepsilon v^2}{2s-2v}, \tag{4.11}$$

$$\begin{aligned} w(s)-2v-3\varepsilon v^2&=\frac{2}{(s-v)}\left(v^2+\varepsilon v^3-s^2-\varepsilon s^3-w(s)(v-s)\right)\\ &=2\left(w(s)-v-s-\varepsilon(v^2+vs+s^2)\right). \end{aligned}$$

And thus

$$0=\varepsilon v^2-(2\varepsilon s)v+w(s)-2s-2\varepsilon s^2. \tag{4.12}$$

So

$$v=s\pm\left(3s^2-(w(s)/\varepsilon)+(2s/\varepsilon)\right)^{1/2}. \tag{4.13}$$

Since $v<s$ is desired we choose the $-$ sign. Now by (4.10)

$$\begin{aligned} w'(s)&=\frac{w(s)-(2v+3\varepsilon v^2)}{s-v}\\ &=2+6\varepsilon s-4\varepsilon\left(3s^2+\frac{2s-w(s)}{\varepsilon}\right)^{1/2}. \end{aligned} \tag{4.14}$$

Now, differentiating (4.13) with respect to s, and substituting (4.14), we compute

$$\frac{dv}{ds}=1-\frac{6s+\dfrac{2-w'(s)}{\varepsilon}}{2\left(3s^2+\dfrac{2s-w(s)}{\varepsilon}\right)^{1/2}}$$

(4.15)

$$=1-\frac{6s+\dfrac{2}{\varepsilon}-\dfrac{1}{\varepsilon}\left(2+6\varepsilon s-4\varepsilon\sqrt{3s^2+\dfrac{2s-w(s)}{\varepsilon}}\right)}{2\sqrt{3s^2+\dfrac{2s-w(s)}{\varepsilon}}}=-1.$$

Thus $v=-s+c$. Since, as depicted in Fig. 14, $v=0$ when $s=u_0$, we have that $c=u_0$. Thus $v=u_0-s$, and we expect that the surfaces $x=f'(s)t$, $y=w(s)t$, and $x=f'(v(s))t$, $y=f_\varepsilon'(v(s))t$ meet when $s=v(s)=u_0-s$; that is, at $s=u_0/2$.

We may now rewrite (4.10):

$$w'(s)=\frac{w(s)-2v-3\varepsilon v^2}{s-v}=\frac{w(s)+2s-2u_0-3\varepsilon(s-u_0)^2}{2s-u_0}. \tag{4.16}$$

Thus we have a linear first order differential equation for w. The initial condition is $w(u_0)=\gamma(u_0)$. The solution is

(4.17)

$$w(s)=\frac{1}{2}\left(s-\frac{u_0}{2}\right)^{1/2}\left[\int_{u_0}^{s}\frac{2z-2u_0-3\varepsilon(u_0-z)^2}{\left(z-\frac{u_0}{2}\right)^{3/2}}dz+\frac{2\gamma(u_0)}{\left(\frac{u_0}{2}\right)^{1/2}}\right]$$

$$=\frac{1}{2}\Bigg[2(2+3\varepsilon u_0)\left(\left(s-\frac{u_0}{2}\right)-\left(\frac{u_0}{2}\right)^{1/2}\left(s-\frac{u_0}{2}\right)^{1/2}\right)$$

$$+2u_0\left(1+3\varepsilon\frac{u_0}{4}\right)\left(1-\left(s-\frac{u_0}{2}\right)^{1/2}\left(\frac{2}{u_0}\right)^{1/2}\right)$$

$$-2\varepsilon\left(\left(s-\frac{u_0}{2}\right)^2-\left(s-\frac{u_0}{2}\right)^{1/2}\left(\frac{u_0}{2}\right)^{3/2}\right)+2\gamma(u_0)\left(\frac{2}{u_0}\right)^{1/2}\left(s-\frac{u_0}{2}\right)^{1/2}\Bigg].$$

Note that

$$w\left(\frac{u_0}{2}\right)=\left(2u_0-3\varepsilon\frac{u_0^2}{2}\right)\Big/2=2\left(\frac{u_0}{2}\right)-3\varepsilon\left(\frac{u_0}{2}\right)^2=f_\varepsilon'\left(\frac{u_0}{2}\right).$$

Thus the two surfaces $x=f'(s)t$, $y=w(s)t$, and $x=f'(v(s))t$, $y=f_\varepsilon'(v(s))t$, do meet at $s=v(s)=u_0/2$. Since by (4.3)

$$\gamma(u_0)=2u_0-2\varepsilon(u_0)^2\ln(u_0)+(\varepsilon-1)(u_0)^2$$

$$=2\exp\left(-\frac{1+\varepsilon}{2\varepsilon}\right)-2\varepsilon\exp\left(-\frac{1+\varepsilon}{2\varepsilon}\right)\left(-\frac{1+\varepsilon}{2\varepsilon}\right)+(\varepsilon-1)\exp\left(-\frac{1+\varepsilon}{\varepsilon}\right) \tag{4.18}$$

$$=2u_0+2\varepsilon u_0^2,$$

we have that

$$w(s)=-\varepsilon\left(s-\frac{u_0}{2}\right)^2+(2+3\varepsilon u_0)\left(s-\frac{u_0}{2}\right)+u_0\left(1+3\varepsilon\frac{u_0}{4}\right). \tag{4.19}$$

Thus

$$\begin{aligned} w'(s)&=-2\varepsilon\left(s-\frac{u_0}{2}\right)+2+3\varepsilon u_0,\\ w''(s)&=-2\varepsilon<0, \quad \text{for } \varepsilon>0, \end{aligned} \tag{4.20}$$

and we see that the shock curve $y=w(s)$, $x=f'(s)=2s$ is concave down. Furthermore, $w'(u_0/2)=2+3\varepsilon u_0=2+6\varepsilon(u_0/2)=f_\varepsilon''(u_0/2)$. Thus the shock curve meets the curve $y=f_\varepsilon'(v)$, $x=f'(v)$ tangentially.

To verify the entropy condition for Ω, it is necessary to show that

$$\begin{aligned} W(k,s)&=(2w(s)-2sw'(s))(k-(u_0-s))\\ &\quad+\left(k^2-(u_0-s)^2\right)w'(s)-\left(k^2+\varepsilon k^3-(u_0-s)^2-\varepsilon(u_0-s)^3\right)2\geq 0, \end{aligned} \tag{4.21}$$

for $u_0-s<k<s$, $(u_0/2)\leq s\leq u_0$. Substituting (4.19), we have

$$\begin{aligned} W(k,s)&=(k+s-u_0)\left(2\varepsilon s^2-2\varepsilon u_0^2\right)+\left(k^2-u_0^2+2u_0s-s^2\right)[4\varepsilon u_0-2\varepsilon s]\\ &\quad-2\left[\varepsilon k^3-\varepsilon u_0^3+3\varepsilon u_0^2 s-3\varepsilon u_0 s^2+\varepsilon s^3\right]. \end{aligned} \tag{4.22}$$

It is now easily checked that $W(s,s)=W(u_0-s,s)=0$; this verifies Condition 2.1, the Rankine–Hugoniot condition, for Ω. Next

$$\begin{aligned} \frac{\partial W}{\partial k}(k,s)&=\left(2\varepsilon s^2-2\varepsilon u_0^2\right)+2k(4\varepsilon u_0-2\varepsilon s)-6\varepsilon k^2\\ &=0 \quad \text{for } k=-\frac{1}{3}s+\frac{2}{3}u_0\pm\frac{1}{3}(2s-u_0)=\frac{1}{3}s+\frac{1}{3}u_0 \quad \text{or} \quad u_0-s. \end{aligned} \tag{4.23}$$

Thus for each s, $W(k,s)$ is cubic in k, with a single root at $k=s$ and a double root at $k=u_0-s=v$. Since these are the only zeros of W, and the leading coefficient of W is $-2\varepsilon<0$, and $v<s$, we conclude that $W\geq 0$ for $v\leq k\leq s$.

Note that in the above proof ε may be arbitrarily large. Furthermore $\lim_{\varepsilon\to\infty}u_0=e^{-.5}\cong .61<1$. Also, the double root of W at $k=v$ shows that Ω is a two dimensional analog of a one dimensional scalar contact discontinuity.

Finally, to prove that $\Gamma[1,0:u_0]$ satisfies the entropy condition for all $\varepsilon>0$, it suffices to show

$$W(k,s)=(2\gamma(s)-2s\gamma'(s))(k-0)+(k^2-0)\gamma'(s)-2(k^2+\varepsilon k^3-0)\geq 0, \tag{4.24}$$

for $0\leq k\leq s$, $u_0=e^{-(1+\varepsilon)/2\varepsilon}\leq s\leq 1$. Substituting (4.3) and (4.4), we have

$$\begin{aligned} W(k,s)&=\big(2(2s-2\varepsilon s^2\ln(s)+(\varepsilon-1)s^2)\\ &\qquad\qquad -2s(2-4\varepsilon s\ln(s)-2s)\big)k\\ &\quad+k^2(2-4\varepsilon s\ln(s)-2s)-2(k^2+\varepsilon k^3)\\ &=2k\left(s^2(2\varepsilon\ln s+\varepsilon+1)\right)+k^2(-4\varepsilon s\ln(s)-2s)-2\varepsilon k^3. \end{aligned} \tag{4.25}$$

Thus $W(k,s)=0$ if $k=0$, or if $k=s$. We have already seen, in (4.5), that $\frac{\partial W}{\partial k}(0,s)>0$ if $s>u_0$, and $\frac{\partial W}{\partial k}(0,u_0)=0$. Since W is cubic in k with leading coefficient $-2\varepsilon<0$, we may conclude that $W\geq 0$ for $0\leq k\leq s$, $u_0\leq s\leq 1$.

Acknowledgment. This paper is an edited version of a doctoral dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Horace H. Rackham School of Graduate Studies at the University of Michigan. This dissertation was written under the direction of Professor Joel A. Smoller, to whom I am deeply grateful for inspiration and guidance.

Microfilm copies of this dissertation are available for study at the Library of Congress and the University of Michigan Library.

REFERENCES

[1] E. CONWAY AND J. SMOLLER, *Global solutions of the Cauchy problem for quasi-linear first order equations in several space variables*, Comm. Pure Appl. Math., 19 (1966), pp. 95–105.

[2] J. GLIMM, *Solutions in the large for nonlinear systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[3] J. GLIMM, D. MARCHESIN AND O. MCBRYAN, *The Buckley–Leverett equation: theory, computation and application*, in Proc. Third Meeting of the International Society for the Interaction of Mechanics and Mathematics, Edinburgh, September 10–13, 1979 (to appear).

[4] J. GLIMM, D. MARCHESIN AND O. MCBRYAN, *Unstable fingers in two phase flow*, Comm. Pure Appl. Math., 34 (1981), pp. 53–75.

[5] J. GUCKENHEIMER, *Shocks and rarefactions in two space dimensions*, Arch. Rational Mech. Anal., 59 (1975), pp. 281–291.

[6] S. N. KRUZKOV, *Generalized solutions of the Cauchy problem in the large for nonlinear equations of first order*, Soviet Math. Dokl., 10 (1969), pp. 785–788.

[7] D. W. PEACEMAN, *Fundamentals of Numerical Reservoir Simulation*, Developments in Petroleum Science, 6, Elsevier, New York, 1977, pp. 1–34.

[8] Y. VAL'KA, *Discontinuous solutions of a multidimensional quasilinear equation (Numerical experiments)*, USSR Comp. Math. and Math. Phys., 8 (1968), pp. 257–264.

[9] A. I. VOL'PERT, *The spaces BV and quasilinear equations*, Math. USSR-Sb., 2 (1967), pp. 225–267.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0012 $01.25/0

NEW RESULTS ON THE VIBRATING STRING WITH A CONTINUOUS OBSTACLE*

A. BAMBERGER† AND M. SCHATZMAN†,‡

Abstract. We give an explicit formula which describes the solution of the problem of the linear elastic string vibrating against a plane obstacle without loss of energy. This formula allows us to prove continuous dependence on the initial data; a regularity result in some bounded variation spaces is given. A numerical scheme is deduced from the explicit formula.

Finally we prove the weak convergence of a subsequence of solutions of the penalized problem to a "weak" solution (i.e. one which does not necessarily conserve energy) of the problem with an obstacle when the obstacle is arbitrary; when the obstacle is plane, all the sequence strongly converges to the solution of the obstacle problem which conserves the energy.

1. Introduction.

1.1. Presentation of the problem and the results. This paper aims to give some new results on vibrating strings with obstacles. The model is the same as in [5], but as it appears necessary to elucidate several points of the modelization which was exposed there, we shall give it from the beginning.

We consider the small transverse vibrations of a string that is constrained to be on one side of a material obstacle. Let the transverse displacement at time t of the material point of the string with coordinate x be denoted by $u(x,t)$. If the string were free, i.e., if there was no obstacle, then u would satisfy the wave equation

$$\Box u \equiv u_{tt} - u_{xx} = 0.$$

We assume that the obstacle has position $\varphi(x)$. We translate the requirement that the string stay on one side of the obstacle into the inequality

$$u(x,t) \geq \varphi(x) \quad \forall x,t. \tag{1}$$

When the string does not touch the obstacle, its motion satisfies the wave equation, and thus

$$\operatorname{supp} \Box u \subset \{(x,t): u(x,t) = \varphi(x)\}. \tag{2}$$

We require that the string does not stick to the obstacle; this can be translated as

$$\Box u \geq 0, \tag{3}$$

which means that the obstacle does not exert a downward force on the string.

Notice that (3) is essentially equivalent to subsonic propagation of interactions. To see this, let $t = \sigma(x)$ be a curve which separates a region $\mathcal{R}$ on the half-plane $\mathbb{R} \times (0, \infty)$ in two open regions $\mathcal{R}^+$ and $\mathcal{R}^-$ where $\Box u$ vanishes. Suppose that $u^+ = u|_{\mathcal{R}^+}$ and $u^- = u|_{\mathcal{R}^-}$ are sufficiently smooth, and that

$$u^{\pm}(x, \sigma(x)) = \varphi(x), \tag{4}$$

$$u^{\pm}(x,t) \geq \varphi(x) \quad \forall (x,t) \in \mathcal{R}. \tag{5}$$

*Received by the editors September 26, 1980, and in final revised form December 17, 1981.

†Centre de Mathématiques Appliquées, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

‡Part of this paper was written when this author was visiting the Universities of Michigan (Ann Arbor), Wisconsin (Madison) and California (Berkeley). The research of this author was sponsored by the U. S. Army under contract DAAG29-75-C-0024, and by the U. S. Air Force Office of Scientific Research under contract C-F49620-79-C-0128.

Then we can compute $\Box u$ in the sense of distributions, with ψ a test function:

$$\langle \Box u, \psi \rangle = -\left\langle \frac{\partial u}{\partial t}, \frac{\partial \psi}{\partial t} \right\rangle + \left\langle \frac{\partial u}{\partial x}, \frac{\partial \psi}{\partial x} \right\rangle$$

$$= \int \left[\left(\frac{\partial u^+}{\partial t} - \frac{\partial u^-}{\partial t} \right)(x, \sigma(x)) + \left(\frac{\partial u^+}{\partial x} - \frac{\partial u^-}{\partial x} \right)(x, \sigma(x)) \sigma'(x) \right] \psi(x, \sigma(x))\, dx. \tag{6}$$

Relation (4) can be differentiated with respect to x, and implies

$$\left(\frac{\partial u^+}{\partial x} - \frac{\partial u^-}{\partial x} \right)(x, \sigma(x)) = -\sigma'(x) \left(\frac{\partial u^+}{\partial t} - \frac{\partial u^-}{\partial t} \right)(x, \sigma(x)). \tag{7}$$

Introducing (7) into (8), we get

$$\langle \Box u, \psi \rangle = \int \left(\frac{\partial u^+}{\partial t} - \frac{\partial u^-}{\partial t} \right)(x, \sigma(x))(1 - \sigma'^2(x)) \psi(x, \sigma(x))\, dx.$$

But hypotheses (4) and (5) ensure that

$$\frac{\partial u^+}{\partial t}(x, \sigma(x)) \geq 0 \quad \text{and} \quad \frac{\partial u^-}{\partial t}(x, \sigma(x)) \leq 0.$$

Therefore, $\Box u$ is nonnegative if and only if $|\sigma'|$ is almost everywhere smaller than 1.

It is not enough to suppose that conditions (1), (2) and (3) are satisfied, as nothing has been said of the evolution of the energy of the string during the collision with the obstacle.

The hypothesis that will be made is that the energy is conserved. This requirement should be analysed from a mathematical point of view as follows: The condition must be local, because the propagation properties of hyperbolic equations suggest it, and it must be satisfied wherever in the x, t half-plane the free wave equation is satisfied. Thus, multiplying by $\partial u / \partial t$ the relation

$$\Box u = 0 \quad \text{on } \mathcal{R}, \tag{8}$$

where $\mathcal{R}$ is an open region such that (8) is satisfied, we obtain a relation in divergence form:

$$\frac{\partial}{\partial t} \left(\left| \frac{\partial u}{\partial t} \right|^2 + \left| \frac{\partial u}{\partial x} \right|^2 \right) - \frac{\partial}{\partial x} \left(2 \frac{\partial u}{\partial t} \frac{\partial u}{\partial x} \right) = 0 \quad \text{in } \mathcal{R}. \tag{9}$$

The operations by which we deduce (9) out of (8) are valid if $\partial u / \partial t$ and $\partial u / \partial x$ are locally square-integrable in $\mathbb{R} \times (0, \infty)$.

The energy condition we shall impose is

$$\frac{\partial}{\partial t} \left(\left| \frac{\partial u}{\partial t} \right|^2 + \left| \frac{\partial u}{\partial x} \right|^2 \right) - \frac{\partial}{\partial x} \left(2 \frac{\partial u}{\partial t} \frac{\partial u}{\partial x} \right) = 0 \tag{10}$$

in the sense of distribution on $\mathbb{R} \times (0, \infty)$.

We could alternatively write it as

$$S_u \overset{\text{def}}{=} \left(-2 u_x u_t, u_x^2 + u_t^2 \right), \qquad \nabla \cdot S_u = 0. \tag{11}$$

Here, the first component of the vector field S_u is the energy density flux, and the second component of the vector field S_u is the energy density.

Notice that (10) *cannot* be deduced by multiplying (3) by $\partial u / \partial t$, as $\partial u / \partial t$ must be expected to be discontinuous on the support of $\Box u$.

For initial conditions such that the free solution corresponding to them is locally of bounded energy, it was proved in [5] that the Cauchy problem (1)–(3) and (11) possesses a unique solution if the function φ is convex.

The approach which led to condition (11) is essentially a mathematical one; from the mechanical point of view, one would like to know if (11) implies that the velocity of the string after collision is the opposite of the velocity of the string before collision. The answer is affirmative, but one has to give a meaning to

$$\frac{\partial u}{\partial t}(x,t+0) = -\frac{\partial u}{\partial t}(x,t-0) \quad \text{if } (x,t) \in \operatorname{supp} \Box u. \tag{12}$$

This was the purpose of [1, part V], where it was shown that if

$$\begin{aligned} &\sigma \text{ is Lipschitz continuous on } \mathbb{R}, \text{ with Lipschitz} \\ &\text{constant 1, and } \sigma \geq 0 \text{ on } \mathbb{R}, \end{aligned} \tag{13}$$

$$\int_{-a}^{a} \left(|u_x(x,t)|^2 + |u_t(x,t)|^2 \right) dx \leq C(a,b) \quad \forall a>0, \quad \forall b>0, \quad \forall t \leq b \tag{14}$$

and if (3) is satisfied, then right and left derivatives can be defined almost everywhere on the noncharacteristic parts of the curve $t=\sigma(x)$.

Moreover, if (11) holds, then for all σ satisfying (13), we have:

$$\left| \frac{\partial^+ u}{\partial t}(x,\sigma(x)) \right| = \left| \frac{\partial^- u}{\partial t}(x,\sigma(x)) \right| \quad \text{a.e. on } \{x : |\sigma'(x)|<1\}. \tag{15}$$

We shall prove in §2 the following explicit formula in the case of the plane obstacle.

Let w be the free solution of the wave equation

$$\begin{aligned} &\Box w = 0, \\ &w(x,0) = u_0(x), \\ &w_t(x,0) = u_1(x). \end{aligned}$$

Let the obstacle be $\varphi = 0$, and let the backward wave cone be

$$T^-_{x,t} \overset{\text{def}}{=} \{(x',t') : 0 \leq t' \leq t - |x-x'|\}.$$

Let us denote by r^- the negative part of a number $r^- = \sup(-r, 0)$. Then the solution of the problem (1)–(3) and (11) is given by

$$u(x,t) = w(x,t) + 2 \sup\{(w(x',t'))^- : (x',t') \in T^-_{x,t}\}.$$

This formula shortens considerably a previous proof [2] of continuous dependence on data, and is the key for the numerical scheme studied in §3. We shall give in §4 a regularity theorem in spaces of bounded variation, in the case of a general concave obstacle.

In §5, we shall consider the functions u_λ which solve the problem

$$\begin{aligned} &\Box u_\lambda - \frac{1}{\lambda}(u_\lambda - \varphi)^- = 0, \\ &u_\lambda(x,0) = u_0(x), \\ &\frac{\partial u_\lambda}{\partial t}(x,0) = u_1(x). \end{aligned} \tag{16}$$

In the first half of this section, we shall prove a weak convergence result, which does not depend on the shape of φ nor on the regularity of the initial data. The limit function will satisfy a set of energy inequalities instead of (11).

In the second half, we shall assume that the obstacle is plane, and that du_0/dx and u_1 are locally of bounded variation. Then the solution of (16) converges strongly in $H^1_{\text{loc}}(\mathbb{R}\times\mathbb{R}^+)$, and its limit is the unique solution of (1)–(3) and (11).

1.2. Notation and summary of previous results. We shall use throughout this paper the following notation and definitions:

V is the set of functions u such that

$$\int_{-a}^{a}\left(|u_x(x,t)|^2+|u_t(x,t)|^2\right)dx\le C(a,b)<+\infty \quad \forall a,b, \quad \forall t<b. \tag{17}$$

w is the free solution of the wave equation:

$$\begin{aligned}&\Box w=0,\\ &w(x,0)=u_0(x),\\ &w_t(x,0)=u_1(x).\end{aligned} \tag{18}$$

If we denote the closure of a set A by $\operatorname{cl}A$, we define a set E by

$$E=\operatorname{cl}\{(x,t):w(x,t)<\phi(x)\}.$$

I is the *domain of influence* defined by

$$I=\cup\{T^+_{x,t}:(x,t)\in E\}, \tag{19}$$

where $T^+_{x,t}$ is the forward wave cone $\{(x',t'):t'\ge t+|x-x'|\}$, and the boundary of I, called the *line of influence*, is given by

$$\begin{aligned}&\partial I=\{(x,t):t=\tau(x)\},\\ &\text{where } \tau \text{ is Lipschitz continuous with Lipschitz constant 1}\end{aligned} \tag{20}$$

(see [5, proposition II.3] for the proof of this claim). The backward wave cone $T^-_{x,t}$ is $\{(x',t'):0\le t'\le t-|x-x'|\}$.

The characteristic coordinates ξ and η are given by

$$\xi=\frac{x+t}{\sqrt{2}}, \qquad \eta=\frac{-x+t}{\sqrt{2}} \tag{21}$$

with the notation $\tilde z(\xi,\eta)=z((\xi-\eta)/\sqrt{2},(\xi+\eta)/\sqrt{2})$ for all functions of two variables x and t.

We shall call problem (P_∞) the following problem:

Given $u_0\in H^1_{\text{loc}}(\mathbb{R})$, $u_1\in L^2_{\text{loc}}(\mathbb{R})$ satisfying the compatibility condition

$$\begin{aligned}&u_0(x)\ge\varphi(x),\\ &u_1(x)\ge 0 \quad \text{a.e. on } \{x:u_0(x)=\varphi(x)\};\end{aligned} \tag{22}$$

find u in V such that

(a) $u\ge\varphi$,

(b) $\operatorname{supp}\Box u\subset\{(x,t):u(x,t)=\varphi(x)\}$,

(c) $\Box u\ge 0$;

$$\frac{\partial}{\partial t}\left(u_x^2+u_t^2\right)+\frac{\partial}{\partial x}(-2u_xu_t)=0 \tag{10}$$

in the sense of distributions in $\mathbb{R}\times\mathbb{R}^+$;

$$\begin{aligned}&u(x,0)=u_0(x),\\ &\frac{\partial u}{\partial t}(x,0)=u_1(x).\end{aligned} \tag{23}$$

The precise statement of the results of existence and uniqueness in [5] is as follows:

THEOREM 0. *Problem* (P_∞) *possesses a unique solution* u *if* φ'' *is nonnegative. Moreover, this solution* u *is the unique solution of the linear problem*

$$
\begin{aligned}
&u\in V,\\
&\Box u_{|\{(x,t):\, t\neq\tau(x)\}}=0,\\
&\frac{\partial u}{\partial t}(x,\tau(x)+0)=-\frac{\partial u}{\partial t}(x,\tau(x)-0)\quad a.e.\ on\{x:\tau(x)>0\ \&\ |\tau'(x)|<1\},\\
&u(x,0)=u_0(x),\\
&\frac{\partial u}{\partial t}(x,0)=u_1(x).
\end{aligned}
\tag{24}
$$

If μ is the measure defined by

$$
\langle\mu,\psi\rangle=-2\int w_t(x,\tau(x))\big(1-\tau'(x)^2\big)\psi(x,\tau(x))\,dx,
\tag{25}
$$

then the solution of (19) is given by the sum of the free solution w and of a convolution

$$
u=w+\mathcal{E}*\mu
\tag{26}
$$

where $\mathcal{E}$ is the elementary solution of the wave equation with support in the positive light cone:

$$
\mathcal{E}=\begin{cases}\dfrac{1}{2} & \text{on } \{(x,t):t\geq|x|\},\\ 0 & \text{elsewhere.}\end{cases}
\tag{27}
$$

It will be useful to consider the problems $(P_{x,t})$, which are just (P_∞) restricted to the backward wave cone $T^-_{x,t}$, with initial data given on $[x-t,x+t]$. Clearly, u is a solution of (P_∞) if and only if it is a solution of $(P_{x,t})$ for all $x\in\mathbb{R}$, $t>0$. The first result on the convergence of the penalty method for the string with an obstacle was proved by A. Bamberger [3].

An explicit formula for the string with a point obstacle was obtained by L. Amerio in [1] and by M. Schatzman in [6], with a different argument.

Continuous dependence on the data and convergence of the penalty method for the point obstacle are proved in [6]. See also the results of C. Citrini [4], where regularity assumptions are relaxed.

2. The explicit formula. Continuous dependence on the initial data.

2.1. The explicit formula for the infinite string. In the case of the zero obstacle (and more generally, the plane obstacle), the solution of (P_∞) can be expressed by an explicit formula. We denote by $r^-=\sup(r,0)$ the negative part of a number.

THEOREM 1. *The unique solution of* (P_∞) *when* $\varphi=0$ *is given by*

$$
u(x,t)=w(x,t)+2\sup_{(x',t')\in T^-_{x,t}}[w(x',t')]^-.
\tag{28}
$$

Remark 2. If the obstacle is plane, i.e., if $\varphi(x)=\alpha x+\beta$, then (28) can be generalized to

$$
u(x,t)=w(x,t)+2\sup_{(x',t')\in T^-_{x,t}}[w(x't')-\varphi(x')]^-.
\tag{29}
$$

To deduce (29) from (28) it is enough to consider $u-\varphi$, and notice that $\Box\varphi=0$.

The proof of Theorem 1 comes in several steps. The first step is the following result:

LEMMA 3. *The set where* $\sup\{[w(x',t')]^- : (x',t') \in T^-_{x,t}\}$ *does not vanish is the interior of the domain of influence I.*

Proof. If $w(x',t')<0$ for some (x',t') in the backward cone $T^-_{x,t}$, then (x,t) belongs to the forward cone $T^+_{x',t'}$, the vertex of which is in the interior of E. Thus (x,t) is in the interior of I. Conversely, if (x,t) belongs to the interior of I, then there exists a point (x',t') in the interior of E such that (x,t) belongs to the interior of $T^+_{x',t'}$. We can choose this (x',t') such that $w(x',t')$ is strictly negative, because the set of (x',t') such that $w(x',t')<0$ is dense in the interior of E. Therefore, $\sup_{(x',t')\in T^-_{x,t}}[w(x',t')]^->0$. □

Let us define

$$k(x,t)=\inf\{w(x',t'):(x',t')\in T^-_{x,t}\}. \tag{30}$$

Then, thanks to Lemma 3, we have, if u is defined by (28),

$$u(x,t)=\begin{cases} w(x,t) & \text{for } t\le\tau(x),\\ w(x,t)-2k(x,t) & \text{for } t\ge\tau(x).\end{cases} \tag{31}$$

LEMMA 4. *Let* u_0 *and* u_1 *satisfy the compatibility conditions* (22), *and let I be nonempty. Then the function k satisfies*

$$\Box k=0 \quad \textit{in the interior of } I. \tag{32}$$

Proof. Let us extend w to the whole plane $\mathbb{R}\times\mathbb{R}$, by solving the (backward) wave equation

$$\begin{aligned} &w(x,0)=u_0(x),\\ &w_t(x,0)=u_1(x),\\ &\Box w=0 \quad \text{for } t<0,\ x\in\mathbb{R}.\end{aligned} \tag{33}$$

The assumption that I is not empty implies that, on the line of influence,

$$\begin{aligned} &w(x,\tau(x))=0 \quad \text{if } |\tau'(x)|<1,\\ &w_t(x,\tau(x))<0 \quad \text{a.e. on } \{x:|\tau'(x)|<1\}.\end{aligned}$$

We shall prove that $w(x,t)\ge 0$ for $t\le\tau(x)$, by essentially the same argument as in [5, Thm. IV.2]. For the convenience of the reader, let us sketch it here.

Let $U=\{x:w(x,\tau(x))>\varphi(x)\}=\cup_i]a_i,b_i[$, where the open sets $]a_i,b_i[$ are the connected components of U. Then [5, Lemma II.6] tells us that

$$\tau(x)=\min(\tau(a_i)+x-a_i,\tau(b_i)+b_i-x) \quad \forall x\in[a_i,b_i]. \tag{34}$$

Therefore if we set

$$\begin{aligned} &\xi_i=\frac{a_i+\tau(a_i)}{\sqrt{2}}, && \xi'_i=\frac{b_i+\tau(b_i)}{\sqrt{2}},\\ &\eta_i=\frac{-a_i+\tau(a_i)}{\sqrt{2}}, && \eta'_i=\frac{-b_i+\tau(b_i)}{\sqrt{2}}\end{aligned} \tag{35}$$

the line of influence in characteristic coordinates is such that

$$Y(\xi)=\begin{cases}\eta_i & \text{if } \xi\in[\xi_i,\xi'_i),\\ [\eta'_i,\eta_i] & \text{if } \xi=\xi'_i,\end{cases} \tag{36}$$

if Y is the multivalued mapping (see Fig. 1) defined by

$$\eta \in Y(\xi) \Leftrightarrow \frac{\xi+\eta}{\sqrt{2}} = \sigma\left(\frac{\xi-\eta}{\sqrt{2}}\right).$$

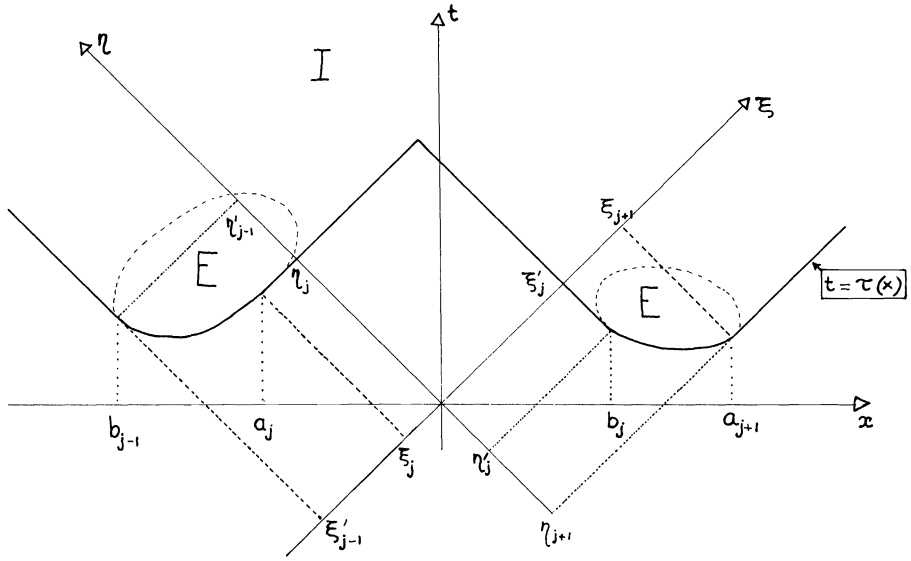


FIG. 1. *The sets E and I, the influence line $t=\tau(x)$, the intervals (a_j, b_j), and the characteristic coordinates, with the ξ_j, ξ_j', η_j, η_j'.*

Let $\tilde{w}(\xi,\eta)=f(\xi)+g(\eta)$, where f and g are in $H^1_{\text{loc}}(\mathbb{R})$. From (35), we deduce $f(\xi)+g(\eta_i')\geq 0$ for $\xi_i\leq\xi\leq\xi_i'$ and from (34), $f(\xi)+g(\eta_i)\geq 0$ for $\xi_i\leq\xi\leq\xi_i'$. As we must have $f(\xi_i)+g(\eta_i)=0=f(\xi_i')+g(\eta_i')=0$, by definition of U, ξ_i, ξ_i', η_i and η_i', then

$$f(\xi_i)=f(\xi_i')\leq f(\xi) \quad \forall \xi\in[\xi_i,\xi_i'] \tag{37}$$

Similarly,

$$g(\eta_i)=g(\eta_i')\leq g(\eta) \quad \forall \eta\in[\eta_i',\eta_i] \tag{38}$$

On C, the complement of the set $\cup_i[\xi_i,\xi_i']$, we have

$$f'(\xi)\leq 0 \quad \text{a.e.} \tag{39}$$

The simplest way to see this is to notice that Y is one-valued on C, and that

$$f(\xi)+g(Y(\xi))=0 \quad \text{on } C,$$
$$f(\xi')+g(Y(\xi))\geq 0 \quad \text{for } \xi'\leq\xi.$$

Let us now evaluate $f(\xi)+g(\eta)$. Suppose first that $X(\eta)=Y^{-1}(\eta)$ is one-valued. Then

$$\begin{aligned} f(\xi)+g(\eta) &= g(\eta)+f(X(\eta))-\int_{\xi}^{X(\eta)} 1_C f'(\xi')\,d\xi' \\ &\quad -\sum_i \left[f(\min(\xi_i', X(\eta)))-f(\max(\xi_i,\xi))\right], \end{aligned} \tag{40}$$

where the summation is extended to the indices such that $[\xi_i, \xi_i']$ intersects $[\xi, X(\eta)]$. We have:

$$g(\eta)+f(X(\eta))\geq 0,$$

$$\int_{\xi}^{X(\eta)} 1_C f'(\xi')\,d\xi' \leq 0 \quad \text{by (39)}.$$

As $X(\eta)$ is one-valued, it is not contained in the interior of an interval $[\xi_i, \xi_i']$. Thus

$$\min[\xi_i', X(\eta))=\xi_i' \quad \text{if } [\xi_i, \xi_i'] \cap [\xi, X(\eta)] \neq \varnothing$$

and, if $\xi \notin [\xi_i, \xi_i']$, the corresponding term in the sum vanishes. For ξ in $[\xi_i, \xi_i']$, the term in the sum is $f(\xi_i')-f(\xi)$, which is not positive, by (37). Therefore, the expression (40) is nonnegative for $\xi \leq X(\eta)$. If we suppose that $X(\eta)=[\xi_j, \xi_j']$, we have to study the expression

$$g(\eta)+f(\xi_j')-\int_{\xi}^{\xi_j'} 1_C f'(\xi')\,d\xi' - \sum_i \left[f\big(\min(\xi_i', \xi_j')\big) - f\big(\max(\xi_i, \xi)\big)\right],$$

and the result still holds, i.e.,

$$w(x,t)\geq 0 \quad \text{for } t \leq \tau(x). \tag{41}$$

Thanks to (41), we may redefine k as

$$k(x,t)=\inf\{w(x',t') : t' \leq t-|x-x'|\},$$

or still, in characteristic coordinates,

$$\tilde{k}(\xi,\eta)=\inf\{f(\xi')+g(\eta') : \xi' \leq \xi \,\&\, \eta' \leq \eta\}. \tag{42}$$

Then, it is immediate that

$$\tilde{k}(\xi,\eta)=\inf\{f(\xi') : \xi' \leq \xi\} + \inf\{g(\eta') : \eta' \leq \eta\}, \tag{43}$$

which proves the claim of Lemma 4. □

We shall now prove that u, defined by (28), satisfies the transmission condition (15) across the line of influence.

LEMMA 5. *If u is defined by* (28), *then almost everywhere on* $\{x : |\tau'(x)|<1\}$

$$\frac{\partial u}{\partial t}(x,\tau(x)+0) = -\frac{\partial u}{\partial t}(x,\tau(x)-0). \tag{44}$$

Proof. Let $A=\{x : |\tau'(x)|<1\}$. Then, almost everywhere on A, by [1, A.2],

$$w_x(x,\tau(x)) \text{ and } w_t(x,\tau(x)) \text{ exist.} \tag{45}$$

Let x be a point satisfying (45), and let us denote

$$w_x(x,\tau(x))=a, \quad w_t(x,\tau(x))=b, \quad \tau'(x)=m.$$

Then

$$a+mb=0, \qquad b\leq 0$$

and

$$w(x',t')=a(x'-x)+b(t'-\tau(x))+\varepsilon(x'-x,t'-\tau(x)),$$

where ε satisfies

$$\lim_{|r|+|s|\to 0} \frac{\varepsilon(r,s)}{|r|+|s|}=0.$$

We have

$$\inf\{a(x'-x)+b(t'-\tau(x)):(x',t')\in T^-_{x,t}\}=b(t-\tau(x)),$$

and therefore,

$$(46)\quad b(t-\tau(x))-\sup\{|\varepsilon(x'-x,t'-t)|:\tau(x')\le t'\le t-|x-x'|\}$$
$$\le k(x,t)\le b(t-\tau(x))+|\varepsilon(0,t-\tau(x))|.$$

As $|\tau'(x)|<1$, we have

$$\lim_{t\downarrow\tau(x)}\left[\sup\{|\varepsilon(x'-x,t'-t)|:\tau(x')\le t'\le t-|x-x'|\}/(t-\tau(x))\right]=0,$$

and we deduce from (46) that

$$\lim_{t\downarrow\tau(x)}\frac{u(x,t)-u(x,\tau(x))}{t-\tau(x)}=-w_t(x,\tau(x))$$

under the assumption (45). □

Conclusion of the proof of Theorem 1. Lemmas 3,4 and 5 imply that the function u defined by (28) solves the linear problem (24), up to the condition $u\in V$. Therefore it remains to check this last condition. If we take into account the formula (43), let us show that k is in V.

We know that f is in $H^1_{\text{loc}}(\mathbb{R})$; let

$$\hat f(\xi)=\inf\{f(\xi'):\xi'\le\xi\}.$$

Then, we can compute the derivative of $\hat f(\xi)$ almost everywhere:

$$(47)\qquad \hat f'(\xi)=\begin{cases}0 & \text{if } f(\xi)>\hat f(\xi) \text{ or if } f'(\xi)\ge 0,\\ f'(\xi) & \text{if } f(\xi)=\hat f(\xi) \text{ and if } f'(\xi)<0.\end{cases}$$

We deduce from (47) that $\hat f$ is in $H^1_{\text{loc}}(\mathbb{R})$. Similarly, $\hat g$ is in $H^1_{\text{loc}}(\mathbb{R})$. The function k which can be written as

$$k(x,t)=\hat f\left(\frac{x+t}{\sqrt{2}}\right)+\hat g\left(\frac{-x+t}{\sqrt{2}}\right)$$

will therefore be in V, i.e.,

$$\int_{-a}^{a}\left(|k_x(x,t)|^2+|k_t(x,t)|^2\right)dx\le C(a,b)\quad\forall a,b,\quad\forall t\ge 0$$

and thus u is in V. □

2.2. Continuous dependence on the data.

COROLLARY 6. *The map* $(u_0,u_1)\to u$ *which to an element of* $H^1_{\text{loc}}(\mathbb{R})\times L^2_{\text{loc}}(\mathbb{R})$ *satisfying the compatibility condition* (22) *associates the solution of* (P_∞) *is continuous from* $H^1_{\text{loc}}(\mathbb{R})\times L^2_{\text{loc}}(\mathbb{R})$ *equipped with the strong topology to*

$$W^{1,p}_{\text{loc}}([0,+\infty);L^2_{\text{loc}}(\mathbb{R}))\cap L^p_{\text{loc}}([0,+\infty);H^1_{\text{loc}}(\mathbb{R}))$$

equipped with the strong topology, for all finite p.

Proof. We have at once the continuity from $H^1_{\text{loc}}(\mathbb{R})\times L^2_{\text{loc}}(\mathbb{R})$ to $C^0(\mathbb{R}\times\mathbb{R}^+)$.

The topology on $W^{1,p}_{\mathrm{loc}}([0,+\infty);\ L^2_{\mathrm{loc}}(\mathbb{R}))\cap L^p_{\mathrm{loc}}([0,+\infty);\ H^1_{\mathrm{loc}}(\mathbb{R}))$ is defined by the seminorms for $A,B>0$

$$q_{ABp}(u)=|u(0,0)|+\left(\int_0^B\left[\int_{-A}^{A}(u_x^2+u_t^2)(x,t)\,dx\right]^{p/2}\right)^{1/p}.$$

The topology on $H^1_{\mathrm{loc}}(\mathbb{R})\times L^2_{\mathrm{loc}}(\mathbb{R})$ is defined by the seminorms for $A>0$

$$p_A(u_0,u_1)=|u_0(0)|+\left(\int_{-A}^{A}\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx\right)^{1/2}.$$

It has been proved in [5, §IV.2] that for solutions of (P_∞) with zero obstacle,

$$\left|\frac{\partial u}{\partial\xi}(x,t)\right|=\left|\frac{\partial u}{\partial\xi}(x+t,0)\right|,\qquad \left|\frac{\partial u}{\partial\eta}(x,t)\right|=\left|\frac{\partial u}{\partial\eta}(x-t,0)\right|. \tag{48}$$

Therefore

$$\begin{aligned}\int_{-A}^{A}\left(|u_x|^2+|u_t|^2\right)(x,t)\,dx&=\int_{-A}^{A}\left(|u_\xi|^2+|u_\eta|^2\right)(x,t)\,dx\\ &\le\int_{-A-t}^{A+t}\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx\quad\forall A,t>0.\end{aligned} \tag{49}$$

Let $q_{AB\infty}$ be the seminorm

$$q_{AB\infty}(v)=v(0,0)+\operatorname*{ess\,sup}_{t\in[0,B]}\left(\int_{-A}^{A}\left(|v_x|^2+|v_t|^2\right)(x,t)\,dx\right)^{1/2}.$$

Then (49) implies

$$q_{AB\infty}(u)\le p_{A+B}(u_0+u_1). \tag{50}$$

If (u_0^n,u_1^n) is a sequence of initial data satisfying the compatibility condition (22) and converging to (u_0,u_1) in $H^1_{\mathrm{loc}}(\mathbb{R})\times L^2_{\mathrm{loc}}(\mathbb{R})$, then, as a consequence of (50),

$$u^n\to u\quad\text{in } H^1_{\mathrm{loc}}(\mathbb{R}\times\mathbb{R}^+)\text{ weakly}, \tag{51}$$

and moreover, (48) implies that

$$\int_{-A}^{A}\left|\frac{\partial u^n}{\partial\xi}(x,t)\right|^2dx\to\int_{-A}^{A}\left|\frac{\partial u}{\partial\xi}(x,t)\right|^2dx\quad\forall t,A>0 \tag{52}$$

$$\int_{-A}^{A}\left|\frac{\partial u^n}{\partial\eta}(x,t)\right|^2dx\to\int_{-A}^{A}\left|\frac{\partial u}{\partial\eta}(x,t)\right|^2dx\quad\forall t,A>0. \tag{53}$$

Gathering (51), (52) and (53), we obtain

$$u^n\to u\quad\text{in } H^1_{\mathrm{loc}}(\mathbb{R}\times\mathbb{R}^+)\text{ strongly}. \tag{54}$$

Thanks to Fubini's theorem, one has from (54)

(55)

$$\left(u_x^n(\cdot,t),u_t^n(\cdot,t)\right)\to\left(u_x(\cdot,t),u_t(\cdot,t)\right)\quad\text{in }\left(L^2_{\mathrm{loc}}(\mathbb{R})\right)^2\text{ strongly, for almost all } t\ge0.$$

The relation (55) together with the estimate

$$q_{AB\infty}(u^n)\le\sup_n p_{A+B}(u_0^n+u_1^n)<+\infty$$

imply that u^n converges to u in the space $W^{1,p}_{\mathrm{loc}}([0,+\infty);\ L^2_{\mathrm{loc}}(\mathbb{R}))\cap L^p_{\mathrm{loc}}([0,+\infty);\ H^1_{\mathrm{loc}}(\mathbb{R}))$. □

Remark 7. The mapping $(u_0,u_1)\to u$ is not continuous to $W^{1,\infty}_{\mathrm{loc}}([0,+\infty);\ L^2_{\mathrm{loc}}(\mathbb{R}))\cap L^\infty_{\mathrm{loc}}([0,+\infty);\ H^1_{\mathrm{loc}}(\mathbb{R}))$ which is the space V defined in (17).

Take for instance the sequence of initial data

$$u_0^n=1,\qquad u_1^n=\frac{n+1}{n}.$$

As these do not depend on x, the solution of P_∞ is

$$u^n(x,t)=\begin{cases}1-\dfrac{n+1}{n}t & \text{if } t<\dfrac{n}{n+1},\\[2mm] \dfrac{n+1}{n}t-1 & \text{if } t\geq\dfrac{n}{n+1},\end{cases}$$

with the limit

$$u(x,t)=\begin{cases}1-t & \text{if } t<1,\\ t-1 & \text{if } t\geq 1.\end{cases}$$

Then we may calculate $q_{AB\infty}(u^n-u)$:

$$q_{AB\infty}(u^n-u)=\begin{cases}0 & \text{if } B<\dfrac{n}{n+1},\\[2mm] 2\sqrt{2A} & \text{if } B>\dfrac{n}{n+1}.\end{cases}$$

Thus if $B>1$, $q_{AB\infty}(u^n-u)$ does not tend to zero as n tends to infinity.

2.3. Application of the explicit formula to the finite string with fixed ends. The explicit formula (29) will allow us to give a simple construction of the solution of (P_f) where (P_f) is the problem of the vibrating string with fixed ends, and obstacle $\varphi=-K<0$. The only modification with respect to (P_∞) we shall require is that u be in the space $L^\infty(0,T;H^1_0(0,L))\cap W^{1,\infty}(0,T;L^2(0,L))$ for all $T>0$.

In fact, u will be in the space

$$L^\infty\big((0,\infty);H^1_0(0,L)\big)\cap W^{1,\infty}\big((0,\infty);L^2(0,L)\big)$$

because we can integrate (10) on any rectangle $[0,L]\times[0,T]$, and we get the energy equality for arbitrary times T:

$$\int_0^L\Big(|u_x(x,T)|^2+|u_t(x,T)|^2\Big)\,dx=\int_0^L\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx. \tag{56}$$

Let us define

$$e=\left(\int_0^L\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx\right)^{1/2}.$$

Then

$$|u(x,t)-u(0,t)|=\left|\int_0^x u_x(x',t)\,dx'\right|\leq e\sqrt{x},$$

and similarly

$$|u(x,t)-u(L,t)|\leq e\sqrt{L-x}.$$

Let $\alpha = K^2/e^2$. Then

$$\forall t \in [0,\infty), \quad \forall x \in [0,\alpha) \cup (L-\alpha, L], \quad u(x,t) > -K, \tag{57}$$

and $\Box u$ cannot be supported in the strips $([0,\alpha)\cup(L-\alpha,L])\times[0,\infty)$ (cf. Fig. 2). Let us extend the initial conditions u_0, u_1 to the interval $[-\alpha, L+\alpha]$ by:

$$u_i(-x) = -u_i(x) \quad \text{if } x \in [-\alpha, 0], \quad i=0,1,$$
$$u_i(L+x) = -u_i(L-x) \quad \text{if } x \in [0,\alpha], \quad i=0,1.$$

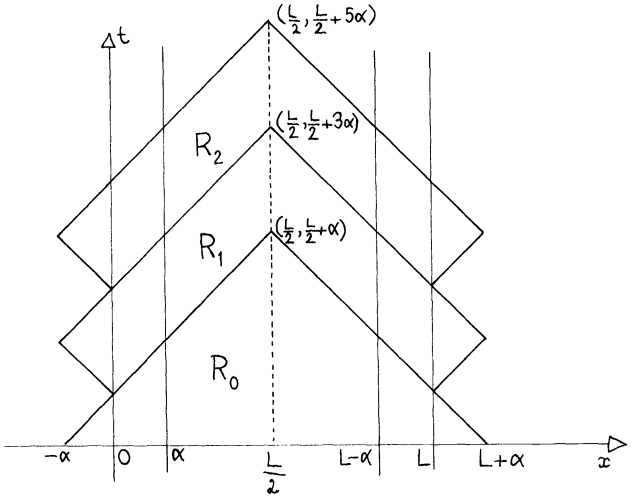


FIG. 2. *The geometric construction used for the explicit formula in the case of the vibrating string with fixed ends, and a constant negative obstacle.*

Then the corresponding free solution w is defined on the cone $T^-_{L/2,L/2+\alpha}$, with the property that

$$w(0,t) = 0, \qquad 0 \le t \le \alpha.$$

Let u be defined on $T^-_{L/2,L/2+\alpha}$ by (29); then for $x=0$, $t \le \alpha$,

$$u(0,t) = w(0,t) + 2 \sup_{T^-_{0,t}} \{[w(x',t') + K]^-\}. \tag{58}$$

But, $w(0,t)=0$, and $T^-_{0,t}$ is included in the strip $[-\alpha,\alpha]\times[0,\infty)$, so that $w > -K$ on this strip, and thus $u(0,t)=0$ on $[0,\alpha]$. Analogously, $u(L,t)=0$ on $[0,\alpha]$.

Therefore, (58) defines the solution of (P_f) on $T^-_{L/2,L/2+\alpha} \cap ([0,L]\times[0,\infty))$.

Let us define by induction the solution of (P_f) on the region R_n given by

$$R_n = \left\{(x,t) \in [0,L]\times[0,\infty): \frac{L}{2} + (2n-1)\alpha - \left|x - \frac{L}{2}\right| \le t \right.$$
$$\left. \le \frac{L}{2} + (2n+1)\alpha - \left|x - \frac{L}{2}\right|\right\}. \tag{59}$$

We shall denote by σ_n the function

$$\sigma_n(x) = \frac{L}{2} + (2n-1)\alpha - \left|x - \frac{L}{2}\right| \quad \text{if } x \in [0,L],$$
$$\sigma_n(x) = \sigma_n(-x), \qquad \sigma_n(L+x) = \sigma_n(L-x) \quad \text{if } x \in [0,\alpha].$$

Suppose we know $u(x,\sigma_n(x))$ for $x\in[0,L]$. Let

$$
\begin{aligned}
&w_n(x,\sigma_n(x))=u(x,\sigma_n(x)) && \text{if } x\in[0,L],\\
&w_n(x,\sigma_n(x))=-u(-x,\sigma_n(-x)) && \text{if } x\in[-\alpha,0],\\
&w_n(x,\sigma_n(x))=-u(2L-x,\sigma_n(2L-x)) && \text{if } x\in[L;L+\alpha],\\
&\Box w_n=0.
\end{aligned}
\tag{60}
$$

The function w_n is defined in the region

$$
\sigma_n(x)\le t\le(2n+1)\alpha+\frac{L}{2}-\left|x-\frac{L}{2}\right|, \qquad -\alpha\le x\le L+\alpha.
$$

Let us notice that the symmetry of the initial conditions in (60) implies

$$
w_n(0,t)=w_n(L,t)=0 \quad \text{for } (2n-1)\alpha\le t\le(2n+1)\alpha. \tag{61}
$$

Moreover, as $u(x,\sigma_n(x))\ge -K$, $\forall x\in[0,L]$, and as u satisfies the energy condition (10), we shall have

$$
\begin{aligned}
&w_n(x,t)\ge -K \quad \text{for } \sigma_n(x)\le t\le(2n+1)\alpha+-|x|\\
&\qquad\qquad \text{or for } \sigma_n(x)\le t\le(2n+1)\alpha+-|L-x|.
\end{aligned}
\tag{62}
$$

Let

$$
u(x,t)=w_n(x,t)+2\sup\left[\left(w_n(x',t')+K\right)^-:\sigma_n(x')\le t'\le t-|x-x'|\right]. \tag{63}
$$

Thanks to (61) and (62), u satisfies the boundary conditions. Therefore it solves the problem of the string with an obstacle on R_n, and the induction can be pursued.

3. A numerical scheme.

3.1. A numerical scheme in a backward cone for the zero obstacle. Let there be given initial data u_0 and u_1 on the interval $[-T,T]$. We seek an approximation to the problem $(P_{0,T})$ on the backward cone $T_{0,T}^-$.

Let $h=T/n$ be a step, and let us define discretized initial data u_0^n and u_1^n by the following formula, where u_0^h is an affine interpolation, and u_1^h is piecewise constant:

$$
\begin{aligned}
&u_0^h(x)=\frac{1}{h}\left[u_0((p+1)h)-u_0(ph)\right](x-ph)+u_0(ph) && \text{if } x\in[ph,(p+1)h],\\
&u_1^h(x)=\frac{1}{h}\int_{ph}^{(p+1)h}u_1(x')\,dx' && \text{if } x\in[ph,(p+1)h].
\end{aligned}
\tag{64}
$$

The corresponding free solution w^h is given by

$$
w^h(x,t)=\frac{1}{2}\left[u_0^h(x+t)+u_0^h(x-t)+\int_{x-t}^{x+t}u_1^h(x')\,dx'\right].
$$

Let us define

$$
\tilde{w}_{i,j}^h=w^h\left(\left(\frac{i-j}{2}\right)h,\left(\frac{i+j}{2}\right)h\right) \quad \text{for } 0\le\frac{i+j}{2}\le n-\left|\frac{i-j}{2}\right|. \tag{65}
$$

Then $\tilde{w}_{i,j}^h$ satisfies the finite difference relation

$$
\tilde{w}_{i,j}^h=\tilde{w}_{i,j-1}^h+\tilde{w}_{i-1,j}^h-\tilde{w}_{i-1,j-1}^h. \tag{66}
$$

Let us define a function $\tilde{u}^h_{i,j}$ on our mesh by

$$\tilde{u}^h_{i,j}=\tilde{w}^h_{i,j}+2\max\left\{\left(\tilde{w}^h_{i',j'}\right)^- : i'\le i, j'\le j,\ i'+j'\ge 0\right\}. \tag{67}$$

We could define $\tilde{u}^h_{i,j}$ alternatively by

$$\tilde{u}^h_{i,j}=\tilde{w}^h_{i,j}-2\tilde{K}^h_{i,j}, \tag{68}$$

$$\begin{aligned}&\tilde{K}^h_{i,j}=\min\left(\tilde{K}^h_{i-1,j},\tilde{K}^h_{i,j-1},-\left(\tilde{w}^h_{i,j}\right)^-\right),\\ &\tilde{K}^h_{i,-i}=0 \quad \text{if } -n\le i\le n.\end{aligned} \tag{69}$$

Notice that $\tilde{K}^h$ is not the discretization of $\tilde{k}$, the correction term in characteristic coordinates, but the discretization of $\tilde{k}.1_{\tilde{I}}$, where $\tilde{I}$ is the set I in characteristic coordinates.

THEOREM 8. *Let u be the solution of* $(P_{0,T})$ *with zero obstacle, and let* $\tilde{u}^h_{i,j}$ *be defined by* (67). *Then*:

$$\max_{i,j}\left|\tilde{u}^h_{i,j}-u\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\le C\sqrt{h}, \tag{70}$$

where C depends only on the initial conditions.

Moreover, we have the following bounds on the (approximate) characteristic derivatives:

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\le\frac{1}{2}\left|\int_{(i-1)h}^{ih}(u_{0x}+u_1)(x')\,dx'\right|, \tag{71}$$

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i,j-1}\right|\le\frac{1}{2}\left|\int_{-jh}^{(-j+1)h}(u_{0x}-u_1)(x')\,dx'\right|. \tag{72}$$

Proof. Let us first evaluate $w^h_{i,j}-w(x',t')$ when (x',t') is in the characteristic square centered on $((i-j)h/2,(i+j)h/2)$, with sides of length $h\sqrt{2}$, i.e.,

$$\frac{2i-1}{2}h\le x'+t'\le\frac{2i+1}{2}h \quad\text{and}\quad \frac{-2j-1}{2}h\le x'-t'\le\frac{-2j+1}{2}h:$$

$$\begin{aligned}w(x',t')-\tilde{w}^h_{i,j}&=w(x',t')-w^h\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\\ &=\frac{1}{2}\Bigg[u_0(x'+t')-u_0^h(ih)+u_0(x'-t')-u_0^h(-jh)\\ &\qquad+\int_{x'-t'}^{x'+t'}u_1(y)\,dy-\int_{-jh}^{ih}u_1^h(y)\,dy\Bigg],\end{aligned}$$

i.e.,

$$\begin{aligned}\left|w(x',t')-\tilde{w}^h_{i,j}\right|&\le\frac{1}{2}\sqrt{\frac{h}{2}}\left[\left|\int_{ih}^{x'+t'}(u_{0x}+u_1)^2dy\right|^{1/2}+\left|\int_{-jh}^{x'-t'}(u_{0x}-u_1)^2dy\right|^{1/2}\right]\\ &\le\frac{1}{2}\sqrt{\frac{h}{2}}\left[2\int_{-T}^{T}\left[(u_{0x}+u_1)^2+(u_{0x}-u_1)^2\right]dy\right]^{1/2}\\ &=\sqrt{\frac{h}{2}}\left(\int_{-T}^{T}\left(u_{0x}^2+u_1^2\right)dx\right)^{1/2}.\end{aligned}$$

We may then deduce from

$$|w(x',t')-\tilde{w}^h_{i,j}|\leq\sqrt{\frac{h}{2}}\left(\int_{-T}^{T}(u_{0x}^2+u_1^2)\,dx\right)^{1/2} \tag{73}$$

that

$$\left|\sup\left\{\left(\tilde{w}^h_{i,j}\right)^-: i'\leq i, j'\leq j, i+j\geq 0\right\}-\sup\left\{[w(x',t')]^-: 0\leq t'<t-|x-x'|\right\}\right| \tag{74}$$
$$\leq\sqrt{\frac{h}{2}}\left(\int_{-T}^{T}(u_{0x}^2+u_1^2)\,dx\right)^{1/2}.$$

Let us note that $\tilde{w}^h_{i,j}=w((i-j)h/2,\ (i+j)h/2)$, because the approximation (64) is very particular.

This, in turn, gives

$$\max_{i,j}\left|\tilde{u}^h_{i,j}-u\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\leq\sqrt{2h}\left(\int_{-T}^{T}(u_{0x}^2+u_1^2)\,dx\right)^{1/2} \tag{75}$$

This completes the proof of (70).

We now turn to proving (71) and (72). Let us note first that if

$$\tilde{k}^h_{i,j}=\min\left\{\tilde{w}^h_{i',j'}:\ i'\leq i, j'\leq j,\ i'+j'\geq 0\right\},$$

we can write $\tilde{k}^h_{i,j}$ alternatively as

$$\tilde{k}^h_{i,j}=\min\left\{\tilde{w}^h_{i',j'}: i'\leq i, j'\leq j\right\}, \tag{76}$$

because we know from (41) that $\tilde{w}^h_{i,j}\geq 0$ for $i+j\leq 0$, as long as we suppose that the domain of influence is not empty.

Relation (76) implies that

$$\tilde{k}^h_{i,j}=\hat{f}^h(i)+\hat{g}^h(j), \tag{77}$$

where

$$\tilde{w}^h_{i,j}=f^h(i)+g^h(j), \tag{78}$$

and

$$\hat{f}^h(i)=\min\{f^h(i'): i'\leq i\},\qquad \hat{g}^h(j)=\min\{g^h(j'): j'\leq j\}. \tag{79}$$

Thus, (68) can be written as

$$\tilde{u}^h_{i,j}=\tilde{w}^h_{i,j}+2\left[\hat{f}^h(i)+\hat{g}^h(j)\right]^-$$

if T is not empty. If $\hat{f}^h(i)+\hat{g}^h(j)\geq 0$, then $\hat{f}^h(i-1)+\hat{g}^h(j)\geq 0$, and (71) is immediate. Suppose now that

$$\hat{f}^h(i)+\hat{g}^h(j)<0. \tag{80}$$

We have two cases. In the first case,

$$\hat{f}^h(i-1)+\hat{g}^h(j)\geq 0. \tag{81}$$

Then, necessarily

$$f^h(i)=\hat{f}^h(i)<\hat{f}^h(i-1)\leq f^h(i-1) \tag{82}$$

and thus,

$$\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}=f^h(i)+g^h(j)-2\hat{f}^h(i)-2\hat{g}^h(j)-f^h(i-1)-g^h(j)$$
$$=-\left[f^h(i)+\hat{g}^h(j)+\hat{f}^h(i-1)+\hat{g}^h(j)\right].$$

Thanks to (80) and (82), we get

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\le\left|f^h(i)+\hat{g}^h(j)\right|+\left|f^h(i-1)+\hat{g}^h(j)\right| \tag{83}$$
$$\le\left|f^h(i)-f^h(i-1)\right|.$$

In the second case,

$$\hat{f}^h(i-1)+\hat{g}^h(j)<0.$$

If $\hat{f}^h(i-1)=\hat{f}^h(i)$, we have immediately

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\le\left|f^h(i)-f^h(i-1)\right|. \tag{84}$$

If $\hat{f}^h(i-1)>\hat{f}^h(i)$, then, we have (82), and

(85)

$$\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}=f^h(i)+g^h(j)-2\hat{f}^h(i)-2\hat{g}^h(j)-f^h(i-1)-g^h(j)+2\hat{f}^h(i-1)+2\hat{g}^h(j)$$
$$=2\hat{f}^h(i-1)-f^h(i)-f^h(i-1),$$

and, thanks to (82) we have

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\le\left|f^h(i)-f^h(i-1)\right|.$$

From (83), (84) and (85), we deduce

$$\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\le\left|w^h_{i,j}-w^h_{i-1,j}\right|=\frac{1}{2}\left|\int_{(i-1)h}^{ih}(u_{0x}+u_1)(x')\,dx'\right|.$$

The proof of (72) is analogous. □

We can deduce from (71) and (72) an energy inequality. Let i_0, j_0 be given such that $-n\le i_0,j_0\le n$ and $i_0+j_0\ge 0$. Then we have

$$\sum_{i=-j_0+1}^{i_0}\frac{1}{h}\left|\tilde{u}^h_{i,j_0}-\tilde{u}^h_{i-1,j_0}\right|+\sum_{j=-i_0+1}^{j_0}\frac{1}{h}\left|\tilde{u}^h_{i_0,j}-\tilde{u}^h_{i_0,j-1}\right|^2 \tag{86}$$
$$\le\frac{1}{2}\int_{-j_0h}^{i_0h}\left[|u_{0x}|^2(x')+|u_1|^2(x')\right]dx'.$$

3.2. A numerical scheme for the string with fixed ends and a constant obstacle. We shall use here the inductive construction of §2.3, which we discretize.

Let u_0 and u_1 be given on $[0,L]$, and let

$$\alpha=K^2\Big/\left[\int_0^L\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx\right], \tag{87}$$

where the obstacle is $\varphi(x)=-K<0$.

Let n be an even integer, and let the step be $h=L/n$; let n_0 be the largest integer such that $n_0h\le\alpha$.

We discretize the initial data as in (64) for $0\le p\le n$, and we extend them as periodic and odd functions:

$$u_r^h(x)=\begin{cases} -u_r^h(-x) & \text{for } -n_0h\le x\le 0, \quad r=0,1,\\ -u_r^h(2L-x) & \text{for } nh\le x\le (n+n_0)h, \quad r=0,1.\end{cases}$$

We define $w^{0,h}$ by

$$\begin{aligned} &w^{0,h}(x,0)=u_0^h(x), \qquad -n_0h\le x\le (n+n_0)h,\\ &\frac{\partial w^{0,h}}{\partial t}(x,0)=u_1^h(x), \qquad -n_0h\le x\le (n+n_0)h,\\ &\Box w^{0,h}=0 \quad \text{in } T^-_{(n/2)h,(n/2+n_0)h}, \end{aligned} \tag{88}$$

and let

$$\tilde w_{i,j}^{0,h}=w^{0,h}\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right). \tag{89}$$

Let

$$\tilde u_{i,j}^h=\tilde w_{i,j}^{0,h}+2\sup\left\{\left(\tilde w_{i,j}^{0,h}+K\right)^-: i'\le i, j'\le j,\ i'+j'\ge 0\right\} \tag{90}$$

where $i\le n+n_0$, $j\le n_0$, $i+j\ge 0$.
Let us define a subset $R^{m,h}$ of $\mathbb{Z}\times\mathbb{Z}$ by

$$\begin{aligned} R^{m,h}=&\left[n+(2m-1)n_0,n+(2m+1)n_0\right]\times\left[-n+(2m-1)n_0,(2m-1)n_0\right]\\ &\cup\left[(2m-1)n_0,n+(2m+1)n_0\right]\times\left[(2m-1)n_0,(2m+1)n_0\right]. \end{aligned} \tag{91}$$

The region $R^{m,h}$ is the discretized equivalent (in i,j coordinates) of the region R^m defined by (59).We define $\tilde w^{m,h}$ on the lower boundary of $R^{m,h}$ by

$$\tilde w_{i,j}^{m,h}=\begin{cases} \tilde u_{i,j}^n & \text{for } i=n+(2m-1)n_0,\ -n+(2m-1)n_0\le j\le (2m-1)n_0,\\ \tilde u_{i,j}^h & \text{for } j=(2m-1)n_0,\ (2m-1)n_0\le i\le n+(2m-1)n_0,\\ \tilde w_{j,i}^{m,h} & \text{for } i=(2m-1)n_0,\ (2m-1)n_0\le j\le (2m+1)n_0,\\ \tilde w_{n+j,-n+i}^{m,h} & \text{for } (2m-1)n_0+n\le i\le (2m+1)n_0+n, j=-n+(2m-1)n,\end{cases} \tag{92}$$

and in $R^{m,h}$, we have

$$\tilde w_{i,j}^{m,h}=\tilde w_{i-1,j}^{m,h}+\tilde w_{i,j-1}^{m,h}-\tilde w_{i-1,j-1}^{m,h} \quad \text{for } (i,j),\ (i-1,j-1) \quad \text{in } R^{m,h}. \tag{93}$$

Then, we shall define $\tilde u_{i,j}^h$ on $R^{m,h}\cap\{(i,j): 0\le (i-j)/2\le n\}$ by

$$\tilde u_{i,j}^h=\tilde w_{i,j}^{m,h}+2\sup\left\{\left(\tilde w_{i',j'}^{m,h}+K\right)^-: i'\le i, j'\le j \text{ and } (i',j')\in R^{m,h}\right\}. \tag{94}$$

Of course (94) is the discretization of (63).

THEOREM 9. *Let u^h be defined by* (93), *and let u be the solution of* (P_f) *on* $[0,L]$ *with obstacle* $-K$. *Then*

$$\max_{i,j}\left|\tilde u_{i,j}^h-u\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\le C^{m+1}\sqrt{h} \tag{95}$$

for (i,j) in the region $R^{m,h}$ defined by (91), *where C depends only on the initial conditions. Moreover, we have the following bounds on the (approximate) characteristic derivatives*:

$$\begin{aligned}
&\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i-1,j}\right|\leq\frac{1}{2}\left|\int_{(i-1)h}^{ih}(u_{0x}+u_1)(x')\,dx'\right|,\\
&\left|\tilde{u}^h_{i,j}-\tilde{u}^h_{i,j-1}\right|\leq\frac{1}{2}\left|\int_{-jh}^{(-j+1)h}(u_{0x}-u_1)(x')\,dx'\right|,
\end{aligned}\tag{96}$$

if u_0 and u_1 are extended to all $\mathbb{R}$ by periodicity and imparity.

Proof. We shall replace the number α defined in (87) by n_0h; for this new value of α, we can perform the construction of the solution of P_f as in 2.3, and we shall compare w^m and $\tilde{w}^{m,h}_{i,j}$ on the regions R^m and $R^{m,h}$.

Thanks to Theorem 8, the relation (95) is verified for $m=0$ and $C\geq(2\int_{-\alpha}^{L+\alpha}(u_{0x}^2+u_1^2)\,dx)^{1/2}$, and the relation (96) is satisfied in R_0.

Suppose that for a certain constant C, (95) and (96) are satisfied in $R^{m-1,h}$.

Then we have

$$\left|\tilde{w}^{m,h}_{i,j}-w^m\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\leq C^m\sqrt{h}\tag{97}$$

for i,j on the lower boundary of $R^{m,h}$ which is the upper boundary of $R^{m-1,h}$.

Then we have

$$\left|\tilde{w}^{m,h}_{i,j}-w^m\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\leq 5C^m\sqrt{h}\quad\text{in } R^{m,h},\tag{98}$$

because $\tilde{w}^{m,h}_{i,j}$ (respectively $w^m(((i-j)/2)h,((i+j)/2)h)$) is the sum of at most five terms $\tilde{w}^{m,h}_{i',j'}$ (respectively $w^m(((i'-j')/2)h,((i'+j')/2)h)$) with i', j' on the lower boundary of $R^{m,h}$ (respectively $(((i'-j')/2)h,((i'+j')/2)h)$) on the lower boundary of R^m).

If we now evaluate the difference $\tilde{w}^{m,h}_{i,j}-w^m(x',t')$ when (x',t') is in the characteristic square centered on $((i-j)/2)h$, $((i+j)/2)h$ with sides of length $h\sqrt{2}$, we have

$$\left|\tilde{w}^{m,h}_{i,j}-w^m(x',t')\right|\leq 5C^m\sqrt{h}+\left|w^m\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)-w^m(x',t')\right|,\tag{99}$$

but we have for P_f the equivalent of (48), i.e.,

$$\begin{aligned}
&\left|\frac{\partial u}{\partial\xi}(x,t)\right|=\left|\frac{\partial u}{\partial\xi}(x+t,0)\right|=\left|\frac{1}{\sqrt{2}}(u_{0x}+u_1)(x-t,0)\right|,\\
&\left|\frac{\partial u}{\partial\eta}(x,t)\right|=\left|\frac{\partial u}{\partial\eta}(x-t,0)\right|=\left|\frac{1}{\sqrt{2}}(u_{0x}-u_1)(x-t,0)\right|,
\end{aligned}$$

if u_0 and u_1 are extended to all of $\mathbb{R}$ by imparity and periodicity.

Therefore

$$\begin{aligned}
&\left|\frac{\partial w^m}{\partial\xi}(x,t)\right|=\frac{1}{\sqrt{2}}\left|(u_{0x}+u_1)(x+t,0)\right|,\\
&\left|\frac{\partial w^m}{\partial\eta}(x,t)\right|=\frac{1}{\sqrt{2}}\left|(u_{0x}-u_1)(x-t,0)\right|.
\end{aligned}\tag{100}$$

Relation (100) allows us to evaluate $w^m(((i-j)/2)h,((i+j)/2)h)-w^m(x',t')$,

$$\left|w^m\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)-w^m(x',t')\right|\leq\sqrt{2h}\left(\int_{-\alpha}^{L+\alpha}\left(u_{0x}^2+u_1^2\right)dx\right)^{1/2}. \tag{101}$$

Let us denote by E the number, which has the dimension of an energy:

$$E=\int_{-\alpha}^{L+\alpha}\left(u_{0x}^2+u_1^2\right)dx.$$

Gathering relations (97), (99) and (101), we obtain:

$$\left|\tilde{u}_{i,j}^h-u\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\leq\left(15C^m+2\sqrt{2E}\right)\sqrt{h}\,.$$

Therefore, if we choose $C=15+2\sqrt{2}\,E$, we have

$$15C^m+2\sqrt{2E}\leq C^{m+1}.$$

The proof of (96) is immediate. □

Remark. For (i,j) in $R^{m,h}$, we have

$$\frac{i+j}{2}\geq(2m-1)n_0,$$

and thus

$$1+m\leq\left(\left(\frac{i+j}{2}h\right)/2n_0h\right)+\frac{3}{2}.$$

Therefore, if $(((i-j)/2)h,((i+j)/2)h)$ converges to (x,t) as h goes to zero, we have from (95):

$$\left|\tilde{u}_{i,j}^h-u\left(\frac{i-j}{2}h,\frac{i+j}{2}h\right)\right|\leq C_1^{t/2\alpha}C_1^{3/2}\sqrt{h}$$

for all $C_1>C$, and for all h small enough.

4. Regularity in spaces of functions of locally bounded variation. This section is dedicated to proving the following result of regularity for an arbitrary concave obstacle φ.

THEOREM 10. *Let u_0 and u_1 be elements of $H^1_{\text{loc}}(\mathbb{R})$ and $L^2_{\text{loc}}(\mathbb{R})$ respectively, such that*

$$\frac{du_0}{dx}\ \textit{and}\ u_1\ \textit{are locally of bounded variation}. \tag{102}$$

Suppose that u_0 and u_1 satisfy the compatibility condition (22), *and that the obstacle is concave.*

Then for all η, the function

$$\xi\to\frac{\partial\tilde{u}}{\partial\xi}(\xi,\eta)$$

defined on $[-\eta,+\infty)$ is locally of bounded variation, and analogously, for all ξ the function

$$\eta\to\frac{\partial\tilde{u}}{\partial\eta}(\xi,\eta)$$

defined on $[-\xi,+\infty)$ is locally of bounded variation.

Proof. We retain the notation of §2.1:

$$U=\{x : w(x,\tau(x))>0\}=\bigcup_i\ ]a_i,b_i[;$$

(35)
$$\xi_i=\frac{a_i+\tau(a_i)}{\sqrt{2}},\qquad \xi_i'=\frac{b_i+\tau(b_i)}{\sqrt{2}},$$
$$\eta_i=\frac{-a_i+\tau(a_i)}{\sqrt{2}},\quad \eta_i'=\frac{-b_i+\tau(b_i)}{\sqrt{2}};$$

(36)
$$Y(\xi)=\begin{cases}\eta_i & \text{if } \xi\in[\xi_i,\xi_i'),\\ [\eta_i',\eta_i] & \text{if } \xi=\xi_i',\end{cases}$$

$$C=\Big(\bigcup_i\ [\xi_i,\xi_i']\Big)^c.$$

We have the following representation of the solution:

(103)
$$\tilde{u}(\xi,\eta)=\begin{cases}f(\xi)+g(\eta) & \text{for } \eta\le Y(\xi),\\ \hat{f}(\xi)+\hat{g}(\eta) & \text{for } \eta\ge Y(\xi),\end{cases}$$

with the transmission conditions:

(104)
$$f(\xi)+g(Y(\xi))=\hat{f}(\xi)+\hat{g}(Y(\xi))=\varphi\left[\frac{\xi-Y(\xi)}{\sqrt{2}}\right],$$

(105)
$$f'(\xi)+g'(Y(\xi))=\left[\hat{f}'(\xi)+\hat{g}'(Y(\xi))\right]$$
if Y is one-valued and $0>Y'(\xi)>-\infty$.

If we differentiate (104) with respect to ξ on C, we get

(106)
$$f'(\xi)+Y'(\xi)g'(Y(\xi))=\hat{f}'(\xi)+Y'(\xi)\hat{g}'(Y(\xi))=\frac{1}{\sqrt{2}}\varphi'\left[\frac{\xi-Y(\xi)}{\sqrt{2}}\right](1-Y'(\xi))$$

(notice that Y is decreasing on C, and therefore almost everywhere differentiable). For $Y'(\xi)=0$, we deduce from (106) that

(107)
$$f'(\xi)=\hat{f}'(\xi)=\frac{1}{\sqrt{2}}\varphi'(\xi-Y(\xi)).$$

For $0>Y'(\xi)>-\infty$, we deduce from (105) and (106) that $\hat{f}'(\xi)+f'(\xi)=\sqrt{2}\,\varphi'(\xi-Y(\xi))$,which contains (107). Therefore, we have

(108)
$$\hat{f}'(\xi)+f'(\xi)=\sqrt{2}\,\varphi'(\xi-Y(\xi))\quad \text{a.e. on } C,$$

and differentiating (104) on C^c,

(109)
$$\hat{f}'(\xi)=f'(\xi)\quad \text{a.e. on } C^c.$$

Let us denote by h the function

(110)
$$h(\xi)=\frac{\partial u}{\partial\xi}(\xi,\eta)$$

where η is fixed throughout the end of this proof, and let $\xi_0=\sup\{\xi'\in X(\eta)\}$. Then, for $\xi\le\xi_0$, $h(\xi)=f(\xi)$ and ξ_0 does not belong to any interval (ξ_i,ξ_i').

To evaluate the total variation of h on a given bounded interval $I=[a,b]$, we have to estimate

$$\begin{aligned} TV(h:I)&=TV(h;I\cap(-\infty,\xi_0))+TV(h;C\cap(\xi_0,+\infty))\\ &\quad+TV(h;C^c\cap(\xi_0,+\infty))+|h(\xi_0+0)-h(\xi_0-0)|\\ &\quad+\sum_{\{i:\xi_0\le\xi_i\le b\}}\left[|h(\xi_i'+0)-h(\xi_i'-0)|+|h(\xi_i+0)-h(\xi_i-0)|\right]. \end{aligned}\tag{111}$$

According to (108) and (109), we have:

$$\begin{aligned} &TV(h;I\cap(-\infty,\xi_0))+TV(h;C\cap(\xi_0,+\infty))+TV(h;C^c\cap(\xi_0,+\infty))\\ &\qquad\le TV(f;I)+TV\left(\sqrt{2}\,\varphi'\left(\frac{\xi-Y(\xi)}{\sqrt{2}}\right)-f(\xi);I\right). \end{aligned}\tag{112}$$

By hypothesis, φ'' is positive, therefore φ' is increasing; as $\xi\mapsto(\xi-Y(\xi))/\sqrt{2}$ is increasing, the right-hand side of (112) is bounded.

The term $|h(\xi_0+0)-h(\xi_0-0)|$ is bounded, because (102) ensures that f, and therefore $\hat{f}$, is locally bounded.

The remaining term in (111) is the sum

$$\sum_{\{i:\xi_0\le\xi_i\le b\}}\left[|h(\xi_i'+0)-h(\xi_i'-0)|+|h(\xi_i+0)-h(\xi_i-0)|\right],\tag{113}$$

which could possibly contain an infinite number of terms. Using (108) and (109), we can write the terms of (113) as

(114)

$$\left|f(\xi_i+0)+f(\xi_i-0)-\sqrt{2}\,\varphi'\left(\frac{\xi_i-\eta_i}{\sqrt{2}}-0\right)\right|+\left|f(\xi_i'+0)+f(\xi_i'-0)-\sqrt{2}\,\varphi'\left(\frac{\xi_i'-\eta_i'}{\sqrt{2}}+0\right)\right|.$$

But we have the following inequalities, deduced from the definition of the line of influence and of the intervals $[\xi_i,\xi_i']$:

$$\begin{aligned} &f(\xi_i-0)-\frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi_i-\eta_i}{\sqrt{2}}-0\right)=a_i^-\le 0,\\ &f(\xi_i+0)-\frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi_i-\eta_i}{\sqrt{2}}+0\right)=a_i^+\ge 0,\\ &f(\xi_i'-0)-\frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi_i'-\eta_i'}{\sqrt{2}}-0\right)=b_i^-\le 0,\\ &f(\xi_i'+0)-\frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi_i'-\eta_i'}{\sqrt{2}}+0\right)=b_i^+\le 0. \end{aligned}\tag{115}$$

We can estimate (114) by

$$|a_i^+ + a_i^-| + |b_i^+ + b_i^-| + \frac{1}{\sqrt{2}}\left[\left|\varphi'\left(\frac{\xi_i - \eta_i}{\sqrt{2}} + 0\right) - \varphi'\left(\frac{\xi_i - \eta_i}{\sqrt{2}} - 0\right)\right|\right. \tag{116}$$
$$\left. + \left|\varphi'\left(\frac{\xi_i' - \eta_i'}{\sqrt{2}} + 0\right) - \varphi'\left(\frac{\xi_i' - \eta_i'}{\sqrt{2}} - 0\right)\right|\right].$$

But

$$|a_i^+ + a_i^-| + |b_i^+ + b_i^-| \le |a_i^+ + a_i^-| + |b_i^+ - a_i^+| + |a_i^+ + a_i^-| + |a_i^- - b_i^-|,$$

and using the sign conditions (115),

$$|a_i^+ + a_i^-| + |b_i^+ + b_i^-| \le 2|a_i^+ + a_i^-| + |b_i^+ - a_i^+| + |b_i^- + a_i^-|$$
$$\le 4TV\left(f(\xi) - \frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi - y(\xi)}{\sqrt{2}}\right); [\xi_i, \xi_i']\right). \tag{117}$$

Carrying (117) and (116) into (113), we obtain:

$$\sum_{\{i\,:\,\xi_0 \le \xi_i \le b\}} \left[|h(\xi_i' + 0) - h(\xi_i' - 0)| + |h(\xi_i + 0) - h(\xi_i - 0)|\right]$$
$$\le 4TV(f; [\xi_0, \xi_0']) + 5TV\left(\frac{1}{\sqrt{2}}\varphi'\left(\frac{\xi - y(\xi)}{\sqrt{2}}\right); [\xi_0, \xi_0']\right). \tag{118}$$

Here, $\xi_0' = \sup\{\xi_i' : \xi_i' \le b\}$. The same argument holds for the other characteristic derivative. The proof of Theorem 10 is complete; notice that we have proved, in fact, that locally, $TV((\partial\tilde{u}/\partial\xi)(\cdot, \eta), I)$ is a bounded function of η, for all bounded I. □

Remark 11. It is not true that under hypothesis (102), $(\partial u/\partial\xi)(\cdot, t)$ or $(\partial u/\partial\xi)(\cdot, t)$ are of bounded variation for all t.

To see it, let us consider the following example. Let

$$w(x,t) = \begin{cases} A - t - a(x+t)^4 \sin\dfrac{1}{x+t} & \text{if } |x+t| \le b, \\ A - t & \text{if } |x+t| \ge b. \end{cases} \tag{119}$$

We choose b such that $\sin(1/b) = 0$, and a such that the curve

$$t = A - a(x+t)^4 \sin\frac{1}{x+t} \tag{120}$$

always has a slope less than 1, for $|x+t| \le b$. For this purpose, we differentiate (120) with respect to x:

$$t' = 4a(1+t')(x+t)^3 \sin\frac{1}{x+t} - a(x+t)^2 \cos\frac{1}{x+t} \cdot (1+t'),$$

and so,

$$|t'| \le a\frac{4b^3 + b^2}{1 - a(4b^3 + b^2)}. \tag{121}$$

Clearly $|t'|$ can be made smaller than 1 if a is sufficiently small.

Then we choose A large enough to have

$$w(x,0)=A-ax^4\sin\frac{1}{x}>0 \quad \text{for } |x|\leq b.$$

Obviously, $du_0/dx=w_x(x,0)$ and $u_1=w_t(x,0)$ are locally of bounded variation.

Thanks to (121), the line of influence is given by (120). We shall now see that $(\partial u/\partial\eta)(\cdot,A)$ is not of bounded variation. The straight line $t=A$ crosses the line of influence infinitely many times, at the points

$$x=\frac{1}{n\pi}-A \quad \text{for } \left|\frac{1}{n\pi}\right|<b, \quad n\in\mathbb{Z},$$

and we have

$$\frac{\partial u}{\partial\eta}(x,A)=\begin{cases} -1 & \text{if } x\in\left(\dfrac{1}{(2k+2)\pi}-A,\dfrac{1}{(2k+1)\pi}-A\right), \quad k>0 \\ & \text{or if } x\in\left(\dfrac{1}{(2k+1)}-A,\dfrac{1}{(2k+2)}-A\right), \quad k<0, \\ +1 & \text{if } x\in\left(\dfrac{1}{(2k+1)\pi}-A,\dfrac{1}{2k\pi}-A\right), \quad k>0 \\ & \text{or if } x\in\left(\dfrac{1}{2k\pi}-A,\dfrac{1}{(2k+1)\pi}-A\right), \quad k<0. \end{cases}$$

This function is not of bounded variation on any interval containing zero.

5. Convergence of the penalty method.

5.1. Weak convergence. This paragraph is dedicated to a general (and unfortunately coarse!) study of the penalized problem

$$\begin{aligned} &\Box u_\lambda-\frac{1}{\lambda}(u_\lambda-\varphi)^-=0, \\ &u_\lambda(x,0)=u_0(x), \\ &\frac{\partial u_\lambda}{\partial t}(x,0)=u_1(x), \end{aligned} \tag{122}$$

where $r^-=\sup(-r,0)$, and φ is an arbitrary continuous function of x, and u_0, u_1 satisfy the compatibility condition (22). The parameter λ is positive, and will tend to zero.

Let us mention that (122) always possesses a unique solution; to see this, it is enough to write (122) in the form of an integral equation, and to use Picard iterations.

PROPOSITION 12. *We have the following estimates for the solution u_λ of* (122):

$$\begin{aligned} &\int_a^b\left[\left|\frac{\partial u_\lambda}{\partial t}(x,\sigma(x))\right|^2+\left|\frac{\partial u_\lambda}{\partial x}(x,\sigma(x))\right|^2+2\left(\frac{\partial u_\lambda}{\partial t}\frac{\partial u_\lambda}{\partial x}\right)(x,\sigma(x))\sigma'(x)\right]dx \\ &\qquad\leq\int_a^b\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx \end{aligned} \tag{123}$$

for all Lipschitz continuous σ with Lipschitz constant 1 *such that $\sigma>0$ on (a,b), $\sigma(a)=\sigma(b)=0$;*

$$\int_{T_{x,t}^-}\frac{1}{\lambda}(u_\lambda(x',t')-\varphi(x'))^-dx'\,dt'\leq C(x,t,u_0,u_1) \tag{124}$$

where C does not depend on λ.

Proof. (i) *Estimate* (123). We have the identity

$$\left(\Box u_\lambda - \frac{1}{\lambda}(u_\lambda - \varphi)^-\right)\frac{\partial u_\lambda}{\partial t}$$

$$= \frac{\partial}{\partial x}\left(-\frac{\partial u_\lambda}{\partial t}\frac{\partial u_\lambda}{\partial x}\right) + \frac{1}{2}\frac{\partial}{\partial t}\left(\left|\frac{\partial u_\lambda}{\partial x}\right|^2 + \left|\frac{\partial u_\lambda}{\partial x}\right|^2 + \frac{1}{\lambda}\left((u_\lambda - \varphi)^-\right)^2\right) = 0.$$

Integrating on the region $\{(x,t): a \le x \le b \text{ and } 0 \le t \le \sigma(x)\}$, we obtain the identity

$$\int_a^b \left[\left|\frac{\partial u_\lambda}{\partial t}(x,\sigma(x))\right|^2 + \left|\frac{\partial u_\lambda}{\partial x}(x,\sigma(x))\right|^2 + 2\sigma'(x)\left(\frac{\partial u_\lambda}{\partial t}\frac{\partial u_\lambda}{\partial x}\right)(x,\sigma(x)) + \frac{1}{\lambda}\left((u_\lambda(x,\sigma(x)) - \varphi(x))^-\right)^2\right] dx$$

$$= \int_a^b \left(\left|\frac{du_0}{dx}\right|^2 + |u_1|^2\right) dx,$$

noting that $(u_\lambda(x,0) - \varphi(x))^- = 0$ for all x. From here, (123) is immediate.

(ii). *Estimate* (124).

We integrate $\Box u_\lambda = (1/\lambda)(u_\lambda - \varphi)^-$ on the backward cone $T^-_{x,t}$:

$$\int_{T^-_{x,t}} \Box u_\lambda \, dx' \, dt'$$

$$= \int_{x-t}^{x+t}\left(\frac{\partial u_\lambda}{\partial t}(x', t - |x - x'|) - u_1(x', 0)\right) dx'$$

$$- \int_0^t \left(\frac{\partial u_\lambda}{\partial x}(x + t - t', t') - \frac{\partial u_\lambda}{\partial x}(x - t + t')\right) dt'$$

$$= \int_{x-t}^{x}\left[\frac{\partial u_\lambda}{\partial t}(x', t - |x - x'|) + \frac{\partial u_\lambda}{\partial x}(x', t - |x - x'|)\right] dx'$$

$$+ \int_x^{x+t}\left[\frac{\partial u_\lambda}{\partial t}(x', t - |x - x'|) - \frac{\partial u_\lambda}{\partial x}(x', t - |x - x'|)\right] dx' - \int_{x-t}^{x+t} u_1(x')\, dx'.$$

Let $\sigma(x') = t - |x - x'|$. Then

$$\int_{T^-_{x,t}} \Box u_\lambda(x', t') \, dx' \, dt'$$

$$\le \int_{x-t}^{x+t}\left[\frac{\partial u_\lambda}{\partial t}(x', \sigma(x')) + \sigma'(x')\frac{\partial u_\lambda}{\partial x}(x', \sigma(x'))\right] dx' + \int_{x-t}^{x+t} |u_1(x')|\, dx',$$

and using the Schwarz inequality and (123), we obtain

$$\int_{T^-_{x,t}} \frac{1}{\lambda}(u_\lambda - \varphi)^- dx' \, dt' \le \left[\int_{x-t}^{x+t}\left[\frac{\partial u_\lambda}{\partial t}(x', \sigma(x')) + \sigma'(x')\frac{\partial u_\lambda}{\partial x}(x', \sigma(x'))\right]^2 dx'\right]^{1/2}\sqrt{2t}$$

$$+ \left(\int_{x-t}^{x+t} |u_1(x')|^2 dx'\right)^{1/2}\sqrt{2t}$$

$$\le 2\sqrt{2t}\left(\int_{x-t}^{x+t}\left(|(u_1(x')|^2 + \left|\frac{du_0}{dx}(x')\right|^2\right) dx'\right)^{1/2}. \qquad \square$$

We need definitions of left and right traces of the characteristic derivatives of a function u.

The following results were proved in [5]: let u be in V (cf. Def. (17)), such that $\Box u$ is a positive measure. Then the function

$$\eta \mapsto \frac{\partial \tilde{u}}{\partial \xi}(\xi,\eta)\Big|_{[a,b]}$$

is increasing from $[-a,\infty)$ to $L^2(a,b)$ for all a,b, and similarly

$$\xi \mapsto \frac{\partial \tilde{u}}{\partial \eta}(\xi,\eta)\Big|_{[c,d]}$$

is increasing from $[-c,\infty)$ to $L^2(c,d)$ for all c,d.

We define

$$\begin{aligned}
\frac{\partial \tilde{u}^r}{\partial \xi}(\xi,\eta) &= \lim_{h\downarrow 0} \frac{\partial \tilde{u}}{\partial \xi}(\xi,\eta+h),\\
\frac{\partial \tilde{u}^l}{\partial \xi}(\xi,\eta) &= \lim_{h\downarrow 0} \frac{\partial \tilde{u}}{\partial \xi}(\xi,\eta-h),\\
\frac{\partial \tilde{u}^r}{\partial \eta}(\xi,\eta) &= \lim_{h\downarrow 0} \frac{\partial \tilde{u}}{\partial \eta}(\xi+h,\eta),\\
\frac{\partial \tilde{u}^l}{\partial \eta}(\xi,\eta) &= \lim_{h\downarrow 0} \frac{\partial \tilde{u}}{\partial \eta}(\xi-h,\eta).
\end{aligned} \tag{125}$$

The functions $\partial \tilde{u}^r/\partial \xi$ and $\partial \tilde{u}^l/\partial \xi$ are defined for all ξ not belonging to the null set N_ξ, and for all η larger than $-\xi$; analogously, the functions $\partial \tilde{u}^r/\partial \eta$ and $\partial \tilde{u}^l/\partial \eta$ are defined for all η not belonging to the null set N_η and for all ξ larger than $-\eta$.

[5, Prop. V.2 and Cor. V.4] tell us that

$$\begin{aligned}
&\frac{\partial u^r}{\partial \xi}(\cdot,\sigma(\cdot)) \in L^2_{\mathrm{loc}}(\mathbb{R};(1+\sigma')\,dx),\\
&\frac{\partial u^r}{\partial \eta}(\cdot,\sigma(\cdot)) \in L^2_{\mathrm{loc}}(\mathbb{R};(1-\sigma')\,dx),\\
&\frac{\partial u^l}{\partial \xi}(\cdot,\sigma(\cdot)) \in L^2_{\mathrm{loc}}(\{x:\sigma(x)>0\},(1+\sigma')\,dx),\\
&\frac{\partial u^l}{\partial \eta}(\cdot,\sigma(\cdot)) \in L^2_{\mathrm{loc}}(\{x:\sigma(x)>0\},(1-\sigma')\,dx).
\end{aligned}$$

Note that the above traces are not continuous functions of u. We have the following example:

$$u_n(x,t) = \begin{cases} 1+\dfrac{1}{n}-t & \text{if } t \le 1+\dfrac{1}{n},\\[2ex] t-\left(1+\dfrac{1}{n}\right) & \text{if } t \ge 1+\dfrac{1}{n}. \end{cases}$$

Then

$$\frac{\partial u^r}{\partial \xi}(x,1) = \frac{1}{\sqrt{2}} \quad \forall x, \qquad \frac{\partial u_n^r}{\partial \xi}(x,1) = -\frac{1}{\sqrt{2}} \quad \forall x, \quad \forall n.$$

We may now state the following result of weak convergence of the penalization:

THEOREM 13. *Given initial conditions* $u_0 \in L^1_{\text{loc}}(\mathbb{R})$ *and* $u_1 \in L^2_{\text{loc}}(\mathbb{R})$ *such that* $u_0 \geq \varphi$ *and* $u_1 \geq 0$ *almost everywhere on the set* $\{x : u_0(x) = \varphi(x)\}$, *there exists a function* u *such that*

$$u \in V, \tag{126}$$

$$u \geq \varphi, \tag{127}$$

$$\Box u \geq 0, \tag{128}$$

$$\operatorname{supp} \Box u \subset \{(x,t) : u(x,t) = \varphi(x)\}, \tag{129}$$

$$\begin{aligned} &\int_a^b \left[\left| \frac{\partial u^r}{\partial \xi}(x,\sigma(x)) \right|^2 (1+\sigma'(x)) + \left| \frac{\partial u^r}{\partial \eta}(x,\sigma(x)) \right|^2 (1-\sigma'(x)) \right] dx \\ &\qquad \leq \int_a^b \left(|u_1|^2 + \left| \frac{du_0}{\partial x} \right|^2 \right) dx, \\ &\int_a^b \left[\left| \frac{\partial u^l}{\partial \xi}(x,\sigma(x)) \right|^2 (1+\sigma'(x)) + \frac{\partial u^l}{\partial \eta}(x,\sigma(x)) \right|^2 (1-\sigma'(x)) \Big] dx \\ &\qquad \leq \int_b^a \left(|u_1|^2 + \left| \frac{du_0}{dx} \right|^2 \right) dx, \end{aligned} \tag{130}$$

for all Lipschitz continuous functions σ, *with Lipschitz constant* 1, *such that* $\sigma(a) = \sigma(b) = 0$, $\sigma > 0$ *on* (a,b),

$$u(x,0) = u_0(x), \tag{131}$$

$$\begin{aligned} &\frac{\partial u}{\partial t}(x,0) = u_1(x) \qquad \text{if } u_0(x) > \varphi(x), \\ &\left| \frac{\partial u}{\partial t}(x,0) \right| \leq u_1(x) \qquad \text{if } u_0(x) = \varphi(x). \end{aligned} \tag{132}$$

Proof. From estimates (123) and (124), we can see that we can extract a subsequence u_μ such that

$$u_\mu \to u \quad \text{weakly} * \text{ in } V. \tag{133}$$

The weak $*$ topology on V is defined by the semi-norms

$$\left| \int u f_1 \right| + \left| \int u_x f_2 \right| + \left| \int u_t f_3 \right|$$

where f_1, f_2 and f_3 are in $L^1(\mathbb{R}^+; L^2(\mathbb{R}))$ with compact support in $\mathbb{R} \times [0, \infty)$. We deduce from (133) that

$$u_\mu \to u \text{ in } C^0(\mathbb{R} \times \mathbb{R}^+) \quad \text{with the compact topology.} \tag{134}$$

Possibly with a new extraction

$$\frac{1}{\mu}(u_\mu - \varphi) \to \nu \quad \text{weakly in } M(\mathbb{R} \times \mathbb{R}^+) \text{ the set of measures on } \mathbb{R} \times \mathbb{R}^+. \tag{135}$$

Therefore

$$\Box u = \nu \geq 0. \tag{136}$$

Relation (123) gives a bound on $((u_\mu-\varphi)^-)^2/\mu$ in L^1_{loc} and thus

$$u\geq\varphi.$$

To check (129), let (x_0,t_0) be a point such that $u(x_0,t_0)>\varphi(x_0)$; thanks to (134) we can find a neighborhood U of (x_0,t_0) and a μ_0 such that $u_\mu(x,t)-\varphi(x)\geq\frac{1}{4}(u(x_0,t_0)-\varphi(x_0))\ \forall\mu<\mu_0$, for all $(x,t)\in U$.

Therefore

$$\Box u_\mu\big|_U=0 \quad \text{for } \mu<\mu_0,$$

and in the limit $\Box u\big|_U=0$. This proves (129).

To prove (130), let σ be given, and ε_0 be a positive number. Let us define for $|\varepsilon|<\varepsilon_0$

$$\sigma_\varepsilon(x)=\begin{cases}\sigma(x)+\varepsilon & \text{if } x\in[a+\varepsilon_0,b-\varepsilon_0],\\ x-a+\varepsilon-\varepsilon_0 & \text{if } x\in[a+\varepsilon_0-\varepsilon,a+\varepsilon_0],\\ -x+b-\varepsilon_0+\varepsilon & \text{if } x\in[b-\varepsilon_0,b+\varepsilon-\varepsilon_0].\end{cases} \tag{137}$$

Then (123) implies

$$\int_{-\varepsilon'}^{\varepsilon''}d\varepsilon\int_{a+\varepsilon_0}^{b-\varepsilon_0}\left[\left|\frac{\partial u_\mu}{\partial\xi}(x,\sigma_\varepsilon(x))\right|^2(1+\sigma_\varepsilon'(x))+\left|\frac{\partial u_\mu}{\partial\eta}(x,\sigma_\varepsilon(x))\right|^2(1-\sigma_\varepsilon'(x))\right]dx \tag{138}$$
$$\leq(\varepsilon'+\varepsilon'')\int_{a-\varepsilon_0}^{b+\varepsilon_0}\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx.$$

But the left-hand side term of (138) can be written as

$$\int_{a+\varepsilon_0}^{b-\varepsilon_0}dx\int_{\sigma(x)-\varepsilon'}^{\sigma(x)+\varepsilon''}\left\{\left|\frac{\partial u_\mu}{\partial\xi}(x,t)\right|^2(1+\sigma'(x))+\left|\frac{\partial u_\mu}{\partial\eta}\right|^2(x,t)(1-\sigma'(x))\right\}dt,$$

and we can take a weak limit in this double integral, thanks to (133).

Thus we can rewrite (138) without the index μ:

$$\int_{-\varepsilon'}^{\varepsilon''}d\varepsilon\int_{a+\varepsilon_0}^{b-\varepsilon_0}\left[\left|\frac{\partial u}{\partial\xi}(x,\sigma_\varepsilon(x))\right|^2(1+\sigma_\varepsilon'(x))+\left|\frac{\partial u}{\partial\eta}(x,\sigma_\varepsilon(x))\right|^2(1-\sigma_\varepsilon'(x))\right]dx \tag{139}$$
$$\leq(\varepsilon'+\varepsilon'')\int_{a-\varepsilon_0}^{b+\varepsilon_0}\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx.$$

Taking $\varepsilon'=0$ in (139) and letting ε'' tend to zero, we obtain

$$\int_{a+\varepsilon_0}^{b-\varepsilon_0}\left[\left|\frac{\partial u^r}{\partial\xi}(x,\sigma(x))\right|^2(1+\sigma'(x))+\left|\frac{\partial u^r}{\partial\eta}(x,\sigma(x))\right|^2(1-\sigma'(x))\right]dx$$
$$\leq\int_{a-\varepsilon_0}^{b+\varepsilon_0}\left(\left|\frac{du_0}{dx}\right|^2+|u_1|^2\right)dx.$$

Letting ε_0 go to zero, we obtain the first relation of (130). If we take $\varepsilon''=0$ and let ε' and then ε_0 tend to zero, we obtain the second relation of (130).

The initial condition (131) is obviously satisfied. It remains to check (132). For this purpose, let us take, in (137), $\sigma(x)=0$ on $[a,b]$. Then, ultimately we get

$$\int_a^b \left[\left| \frac{\partial u^r}{\partial \xi}(x,0) \right|^2 + \left| \frac{\partial u^r}{\partial \eta}(x,0) \right|^2 \right] dx \leq \int_a^b \left(\left| \frac{du_0}{dx} \right|^2 + |u_1|^2 \right) dx.$$

Using the identity

$$\left| \frac{\partial u^r}{\partial \xi}(x,0) \right|^2 + \left| \frac{\partial u^r}{\partial \eta}(x,0) \right|^2 = \left| \frac{du_0}{dx} \right|^2 + \left| \frac{\partial u}{\partial t}(x,0+0) \right|^2,$$

which takes into account (131), we have

$$\int_b^a \left| \frac{\partial u}{\partial t}(x,0+0) \right|^2 dx \leq \int_a^b |u_1|^2 dx.$$

As a and b are arbitrary, we have eventually

$$\left| \frac{\partial u}{\partial t}(x,0+0) \right| \leq |u_1(x)| \quad \text{a.e. on } \mathbb{R}.$$

When $u_0(x) > \varphi(x)$, we have the first part of (132), as locally, $\nu = \Box u = 0$. □

We shall now study the relation between the strong convergence of $\partial u_\lambda/\partial x$ and $\partial u_\lambda/\partial t$, and the verification of the energy condition (11).

LEMMA 14. *Let u_λ be a sequence of solutions of* (122), *converging weakly* to a solution u of* (126)–(132). *Then, u satisfies the energy condition* (11) *if and only if $\partial u_\lambda/\partial t$ and $\partial u_\lambda/\partial x$ converge to $\partial u/\partial t$ and $\partial u/\partial x$ respectively, strongly in $L^2_{\text{loc}}(\mathbb{R}\times[0,\infty))$.*

Proof. Notice first that as $((u_\mu-\varphi)^-/\mu)\cdot 1_K$ converges to $\nu\cdot 1_K$ in $M(\mathbb{R}\times\mathbb{R}^+)$ weakly, for all compact K, and as $(u_\mu-\varphi)^-$ converges to zero uniformly on compact sets, then

$$\int_K \left[\frac{\left((u_\mu-\varphi)^-\right)^2}{\mu} \right] dx'\,dt' \to 0 \tag{140}$$

for any compact set K.

Let $\sigma_h(x') = h - |x-x'|$. Then we have the identity, for any function v,

$$\begin{aligned} &\int_0^t dh \int_{x-h}^{x+h} \left[\left| \frac{\partial v}{\partial \xi}(x,\sigma_h(x)) \right|^2 (1+\sigma_h'(x)) + \left| \frac{\partial v}{\partial \eta}(x,\sigma_h(x)) \right|^2 (1-\sigma_h'(x)) \right] dx \\ &= \int_{x-t}^{x} dx' \int_0^{t+x-x'} 2 \left| \frac{\partial v}{\partial \xi}(x',t') \right|^2 dt' + \int_x^{x+t} dx' \int_0^{t-x+x'} 2 \left| \frac{\partial v}{\partial \eta}(x',t') \right|^2 dt'. \end{aligned} \tag{141}$$

If the limit of the sequence u_μ satisfies (11), then the value of (141) for $v=u$ is

$$\int_0^t dh \int_{x-h}^{x+h} \left(\left| \frac{du_0}{dx} \right|^2 + |u_1|^2 \right) dx. \tag{142}$$

The value of (141) for $v=u_\mu$ is

$$\int_0^t dh \int_{x-h}^{x+h} \left(\left| \frac{du_0}{dx} \right|^2 + |u_1|^2 \right) dx - \int_{0\leq t'\leq t-|x-x'|} \frac{1}{\mu} \left((u_\mu-\varphi)^- \right)^2 dx'\,dt. \tag{143}$$

And according to (140), the limit of (143) is (142). Therefore, as $\partial u_\mu/\partial\xi$ (resp. $\partial u_\mu/\partial\eta$) converges weakly to $\partial u/\partial\xi$ (resp. $\partial u/\partial\eta$) in $L^2_{\text{loc}}([0,\infty)\times\mathbb{R}^+)$, and

$$\lim_{\mu\to 0}\int_{A_{x,t}}\left|\frac{\partial u_\mu}{\partial\xi}\right|^2 dx'\,dt'+\int_{B_{x,t}}\left|\frac{\partial u_\mu}{\partial\eta}\right|^2 dx'\,dt'=\int_{A_{x,t}}\left|\frac{\partial u}{\partial\xi}\right|^2 dx'\,dt'+\int_{B_{x,t}}\left|\frac{\partial u}{\partial\eta}\right|^2 dx'\,dt'$$

where $A_{x,t}=\{(x',t')\in T^-_{x,t}/x'\le 0\}$, $B_{x,t}=T^-_{x,t}\setminus A_{x,t}$, we can conclude that the convergence of $\partial u_\mu/\partial\xi$ and $\partial u_\mu/\partial\eta$ to $\partial u/\partial\xi$ and $\partial u/\partial\eta$ is strong.

Conversely, if $\partial u_\mu/\partial\xi$ (resp. $\partial u_\mu/\partial\eta$) converges strongly to $\partial u/\partial\xi$ (resp. $\partial u/\partial\eta$), then it is straightforward to pass to the limit in (11). □

5.2. Strong convergence when the obstacle is zero and the initial characteristic derivatives are of bounded variation. The first step in this study is to notice that if w is an *affine function*, then the penalized solution converges to the solution of (P_∞) which conserves the energy.

LEMMA 15. *Let there be given initial conditions*

$$\begin{aligned} &u(x,0)=a-bx\ge 0 \quad \text{on } [x_0-t_0,x_0+t_0],\\ &u_t(b,0)=-c<0, \end{aligned} \tag{144}$$

and suppose that the free solution $w(x,t)=a-bx-ct$ is such that

$$w(x_0,t_0)<0.$$

Then the solution u_λ of (122) *with initial conditions* (144) *is given by*

$$u_\lambda(x,t)=\begin{cases} a-bx-ct & \text{for } bx+ct<a,\\ \sqrt{\lambda(c^2-b^2)}\,\sin\dfrac{ct+bx-a}{\sqrt{\lambda(c^2-b^2)}} & \text{for } a\le bx+ct\le a+\pi\sqrt{\lambda(c^2-b^2)},\\ bx+ct-a-\pi\sqrt{\lambda(c^2-b^2)} & \text{for } bx+ct\ge a+\pi\sqrt{\lambda(c^2-b^2)}. \end{cases} \tag{145}$$

Therefore u_λ converges strongly in $H^1(T^-_{x_0,t_0})$ to the solution of (P_{x_0,t_0}).

Proof. Let us compute the solution of (P_{x_0,t_0}):

$$E=\{(x,t)\in T^-_{x_0,t_0}: a-bx-ct<0\}.$$

We see at once that the slope of the line $a=bx+ct$ is smaller than 1, in absolute value. Therefore $I=E$, and

$$u(x,t)=\begin{cases} a-bx-ct & \text{if } a-bx-ct\ge 0,\\ bx+ct-a & \text{if } a-bx-ct\le 0. \end{cases} \tag{146}$$

Let us look for the solution of (122) with initial conditions (144) under the form

$$u_\lambda(x,t)=f_\lambda(bx+ct).$$

Then f_λ must satisfy the ordinary differential equation

$$(c^2-b^2)f''-\frac{1}{\lambda}f^-=0$$

with the initial conditions

$$f_\lambda(a)=0, \qquad f'_\lambda(a)=-1.$$

This problem can be solved immediately and gives (145). Clearly the limit of the sequence u_λ is u, and Lemma 14 allows us to conclude the proof. □

Remark 16. Suppose we replace the function r^- by a function ψ such that

$$\psi(x)=0 \quad \text{if } x\geq 0,$$
$$\psi(x)>0 \quad \text{if } x<0,$$
$$\psi \text{ is continuous, strictly decreasing on } (-\infty,0),$$
$$\psi(-\infty)=\infty.$$

Then the penalized problem

$$\Box \hat{u}_\lambda - \frac{1}{\lambda}\psi(\hat{u}_\lambda - \varphi)=0,$$
$$\hat{u}_\lambda(x,0)=u_0(x),$$
$$\frac{\partial \hat{u}_\lambda}{\partial t}(x,0)=u_1(x)$$

can be studied as above; we get Theorem 14 with almost no change in the proof. Moreover, a phase plane analysis shows easily that in the case of initial data (144) the limit of u_λ is the function (146). We chose the specific penalization (122) because of its simplicity. We need an integral solution of the linear Klein–Gordon equation with initial values given on a curve $t=\sigma(x)$. This is the object of the next lemma.

LEMMA 17. *Let w be a solution of the wave equation on the set $S=\{(x,t)/\sigma(x)\leq t\leq t_0-|x-x_0|\}$ where σ is a Lipschitz continuous function with Lipschitz constant* 1.

Then the unique solution on S of the problem

$$\Box u+\frac{1}{\lambda}u=0,$$
$$u(x,\sigma(x))=w(x,\sigma(x)), \tag{147}$$
$$\frac{\partial u}{\partial t}(x,\sigma(x))=\frac{\partial w}{\partial t}(x,\sigma(x)) \quad a.e.\ on\ \{x:|\sigma'(x)|<1\}$$

is given by

$$u(x,t)=w(x,t)-\frac{1}{2}\int_{S\cap T^-_{x,t}} \mathcal{J}_0\left(\frac{\sqrt{(t-t')^2-(x-x')^2}}{\sqrt{\lambda}}\right) w(x',t')\,dx'\,dt', \tag{148}$$

where

$$\mathcal{J}_0(y)=\sum_{n\geq 0}\frac{(-1)^n}{n!}\left(\frac{y}{2}\right)^{2n} \tag{149}$$

is the Bessel function $\mathcal{J}_0$.

Proof. We verify that if w is a solution of the wave equation in the whole plane and if

$$\bar{w}(x,t)=\begin{cases}0 & \text{if } t<\sigma(x),\\ w(x,t) & \text{if } t\geq\sigma(x),\end{cases}$$

then

$$\langle \Box \bar{w},\varphi\rangle = -\int w(x,\sigma(x))(\varphi_t(x,\sigma(x))+\sigma'(x)\varphi_x(x,\sigma(x)))\,dx$$
$$+\int [w_t(x,\sigma(x))+\sigma'(x)w_x(x,\sigma(x))]\varphi(x,\sigma(x))\,dx.$$

Solving (147) amounts to finding a solution of

$$\left(\Box u+\frac{1}{\lambda}u\right)\Big|_{\{(x,t)\,:\,t>\sigma(x)\}}=0,$$
$$u(x,t)\big|_{\{(x,t)\,:\,t\le\sigma(x)\}}=0,$$
$$u(x,\sigma(x))=w(x,\sigma(x)),$$
$$\frac{\partial u}{\partial t}(x,\sigma(x))=\frac{\partial w}{\partial t}(x,\sigma(x)) \quad \text{a.e. on } \{x:|\sigma'(x)|<1\},$$

which can be written as

$$u=\bar{w}-\frac{1}{\lambda}\mathcal{E}*u \tag{150}$$

where $\mathcal{E}$ is the elementary solution of the wave equation defined by

$$\mathcal{E}(x,t)=\begin{cases}\frac{1}{2} & \text{if } t\ge x,\\ 0 & \text{elsewhere.}\end{cases}$$

The convolution equation (150) has a unique solution given by

$$u=\sum_{n=0}^{\infty}\frac{(-1)^n}{\lambda^n}\left(\overset{n}{\underset{k=1}{*}}\mathcal{E}\right)*\bar{w}. \tag{151}$$

By a simple inductive calculation in characteristic coordinates, we obtain:

$$\left(\overset{n}{\underset{k=1}{*}}\tilde{\mathcal{E}}\right)(\xi,\eta)=\tilde{\mathcal{E}}\cdot\left(\frac{\xi\eta}{2}\right)^{n-1}\frac{1}{((n-1)!)^2}. \tag{152}$$

Therefore

$$\sum_{n=1}^{\infty}\frac{(-1)^n}{\lambda^n}\left(\overset{n}{\underset{k=1}{*}}\mathcal{E}\right)(x,t)=-\frac{1}{2\lambda}\mathcal{E}\cdot\mathcal{J}_0\left(\frac{\sqrt{t^2-x^2}}{\lambda}\right). \tag{153}$$

Together with (153), formula (151) gives (149). □

We can now state the theorem of convergence for penalized solutions:

THEOREM 18. *Let u_0 and u_1 be such that*

$$\frac{du_0}{dx} \text{ and } u_1 \text{ are locally of bounded variation} \tag{154}$$

and suppose that they satisfy the compatibility condition (122). *Then the solution u_λ of* (122) *converges to the solution of* (P_∞) *when λ goes to zero.*

Proof. Let us first notice that on I^c, the complement of the domain of influence, we have, if u is the solution of (P):

$$\Box u=0, \qquad u\ge 0 \quad \text{on } I^c.$$

Therefore

$$u_\lambda=u \quad \text{on } \overline{I^c}$$

and, in particular,

$$
\begin{aligned}
&u_\lambda(x,\tau(x))=u(x,\tau(x))=w(x,\tau(x)),\\
&\frac{\partial u_\lambda}{\partial t}(x,\tau(x))=\frac{\partial u}{\partial t}(x,\tau(x)-0)=\frac{\partial w}{\partial t}(x,\tau(x)) \quad \text{a.e. on } \{x:|\tau'(x)|<1\},
\end{aligned}
\tag{155}
$$

where we recall that τ, the line of influence, is Lipschitz continuous, with Lipschitz constant 1.

We shall now use assumption (154) to obtain more information about the line of influence. We need the following notation (see Fig. 3):

$$
\begin{aligned}
Q_1&=\{(x,t):x\geq|t|\},\\
Q_2&=\{(x,t):t\leq|x|\},\\
Q_3&=\{(x,t):x\leq-|t|\},\\
Q_4&=\{(x,t):t\leq-|x|\}.
\end{aligned}
\tag{156}
$$

We shall denote

$$
\frac{\partial w}{\partial t}(x,t:Q_i)=\lim_{\substack{(h,k)\to 0\\(h,k)\in Q_i}}\frac{\partial w}{\partial t}(x+h,t+k).
\tag{157}
$$

Thanks to (154), $(\partial w/\partial t)(x,t;Q_i)$ is defined for $1\leq i\leq 4$, and we have the formula

$$
\frac{\partial w}{\partial t}(x,t;Q_1)=\frac{1}{\sqrt{2}}\left[\frac{\partial w^l}{\partial\xi}(x,t)+\frac{\partial w^r}{\partial\eta}(x,t)\right],
\tag{158}
$$

with notation as in (125). We have analogous formulae for the three other limits.

LEMMA 19. *Let x be such that $\tau'(x)$ is defined and $|\tau'(x)|<1$. Suppose that*

$$
\max_{1\leq i\leq 4}\frac{\partial w}{\partial t}(x,t;Q_i)<0.
\tag{159}
$$

Then there exists a neighborhood $(x-\varepsilon,x+\varepsilon)$ of x such that $|x'-x|<\varepsilon\Rightarrow\tau'(x')$ has left and right limits at every point and $|\tau'(x'\pm 0)|<1$; moreover

$$
\sup_{1\leq i\leq 4}\frac{\partial w}{\partial t}(x',\tau(x');Q_i)\leq -l<0.
$$

Proof. The hypothesis (159) implies that, in a neighborhood N of $(x,\tau(x))$

$$
\sup_{1<i<4}\frac{\partial w}{\partial t}(x',t';Q_i)\leq -l<0;
$$

therefore $w(x',\cdot)$ is strictly decreasing for x' close enough to x, and moreover, if k is so chosen that $w_t(x,\tau(x))+kw_x(x,\tau(x))<0$, then

$$
w(x+kh,\tau(x)+h)<0.
$$

Thus, there exists a unique solution to the problem

$$
\begin{aligned}
&w(x',\sigma(x'))=0,\\
&\max\big(|x-x'|,|\sigma(x')-\tau(x)|\big)\leq\alpha, \quad \text{where } \alpha \text{ is a small positive number.}
\end{aligned}
\tag{160}
$$

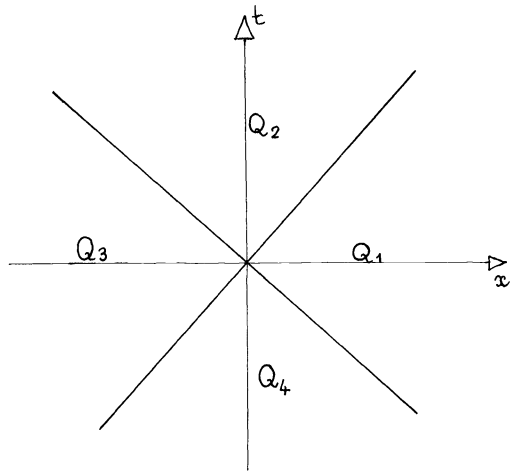


FIG. 3. *The regions* Q_i, $i=1,\cdots,4$ *of the* (x,t)*-plane.*

To prove that σ is identical to τ in an interval $[x-\alpha',x+\alpha']$ where α' may be smaller than α, we have to check that

$$|\sigma'(x')|<1 \quad \text{a.e. on } [x-\alpha',x+\alpha'].$$

The function σ is continuous indeed, as w is continuous and $t'=\sigma(x')$ is the unique solution of $w(x',t')=0$ in N. We may not directly differentiate the relation $w(x',\sigma(x'))=0$, as we do not have the assumptions of the implicit function theorem. But, with the very same argument as in this theorem, and using notation (157) and its analogue for $\partial u/\partial x$, we have

$$\begin{aligned} w(x'+h,\sigma(x+h)) &= w(x,\sigma(x')) + w_x(x',\sigma(x');Q_1)h \\ &\quad + w_t(x',\sigma(x');Q_1)(\sigma(x'+h)-\sigma(x')) \\ &\quad + \varepsilon_1(|h|+|\sigma(x'+h)-\sigma(x')|) \\ &\qquad \text{for all } h \text{ such that } (h,\sigma(x'+h)-\sigma(x'))\in Q_1. \end{aligned} \tag{161}$$

Here ε_1 is a function such that

$$\lim_{k\to 0}\frac{\varepsilon_1(k)}{k}=0.$$

By a standard argument

$$\lim_{\substack{h\to 0 \\ (h,\sigma(x'+h)-\sigma(x'))\in Q_1}}\left[\frac{\sigma(x'+h)-\sigma(x')}{h}\right]=-\frac{w_x(x',\sigma(x');Q_1)}{w_t(x',\sigma(x');Q_1)}. \tag{162}$$

The same result holds in the three other quadrants Q_2, Q_3 and Q_4, and by choosing α' adequately small we shall have

$$\left|\frac{w_x(x',\sigma(x');Q_i)}{w_t(x',\sigma(x');Q_i)}\right|\le 1-\varepsilon \quad \text{for } |x-x'|\le\alpha',$$

and thus

$$(h,\sigma(x'+h)-\sigma(x'))\in Q_1 \quad \text{for } h>0 \text{ small enough,}$$
$$(h,\sigma(x'+h)-\sigma(x'))\in Q_3 \quad \text{for } h<0,\ |h| \text{ small enough,}$$
$$|\sigma'(x')|\le 1-\varepsilon \quad \text{a.e. on } [x-\alpha',x+\alpha'],$$

$\tau(x')=\sigma(x')$, and τ' has right and left limits at all points of $[x-\alpha',x+\alpha']$. □

Let us compare locally the solution of the linear Klein–Gordon equation (147) to the solution of an approaching problem with simpler initial data. Let

$$\tau(x_0)=t_0, \quad \tau_0'(x_0)=m=-\frac{w_x(x_0,t_0)}{w_t(x_0,t_0)},$$
$$\tau_0(x)=t_0+m(x-x_0),$$
$$w_0(x,t)=w_t(x_0,t_0)(t-t_0)+w_x(x_0,t_0)(x-x_0),$$
$$u_0(x,\tau_0(x))=w_0(x,\tau_0(x))=0,$$
$$\frac{\partial u_0}{\partial t}(x,\tau_0(x))=\frac{\partial w_0}{\partial t}(x,\tau_0(x))=w_t(x_0,t_0),$$
$$S_0=\{(x,t):t>\tau_0(x)\}.$$

Then:

$$u_0(x,t)=\sqrt{\lambda(1-m^2)}\,w_t(x_0,t_0)\sin\frac{t-t_0-m(x-x_0)}{\sqrt{\lambda(1-m^2)}}.$$

With the help of (148), we have

$$u_\lambda(x,t)-u_0(x,t)=w(x,t)-w_0(x,t)$$
$$-\frac{1}{2\lambda}\int_{T_{x,t}^-}\mathcal{J}_0\left(\frac{\sqrt{(t-t')^2-(x-x')^2}}{\lambda}\right)\cdot\left[(w\cdot 1_S)(x',t')-(w_0\cdot 1_{S_0})(x',t')\right]dx'\,dt'. \tag{163}$$

Let us estimate (163) for x and t such that

$$|x-x_0|+|t-t_0|\le C\sqrt{\lambda}, \tag{164}$$

and under the hypotheses that $|\tau'(x_0)|<1$ and that $w_t(x_0,t_0)$ and $w_x(x_0,t_0)$ are well defined. Then

$$|w(x,t)-w_0(x,t)|\le o\big(|x-x_0|+|t-t_0|\big)=o\big(\sqrt{\lambda}\big).$$

To estimate the integral, let us first note that

$$|w\cdot 1_S-w_0\cdot 1_{S_0}|\le|w-w_0|\cdot 1_{S\cup S_0}.$$

This relation comes from the fact that, locally, $w\cdot 1_S{}'=-w^-$ and $w_0\cdot 1_{S_0}=-w_0^-$. We define new variables X and T by

$$t-t'=T\sqrt{\lambda}, \qquad x-x'=X\sqrt{\lambda}.$$

Then the integral expression in (163) is estimated by

$$\lambda \int_{T_{0,0}^+} \left| \mathcal{G}_0\left(\sqrt{T^2-X^2}\right)\right| \left| w\left(x-X\sqrt{\lambda}, t-T\sqrt{\lambda}\right) - w_0\left(x-X\sqrt{\lambda}, t-T\sqrt{\lambda}\right)\right| \cdot 1_{S\cup S_0}\left(x-X\sqrt{\lambda}, t-T\sqrt{\lambda}\right) dX\,dT.$$

But

$$\left|w\left(x-X\sqrt{\lambda}, t-T\sqrt{\lambda}\right) - w_0\left(x-X\sqrt{\lambda}, t-T\sqrt{\lambda}\right)\right| \le o\left(\left|x-X\sqrt{\lambda}-x_0\right| + \left|t-T\sqrt{\lambda}-t_0\right|\right),$$

and we have to check that $\{(X,T)\in T_{00}^+ :\ (x-X\sqrt{\lambda}, t-T\sqrt{\lambda})\in S\cup S_0\}$ is bounded. This set can be written as

$$\left\{(X,T): |X| \le T \le \frac{t}{\sqrt{\lambda}} - \min\left(\frac{\tau\left(x-X\sqrt{\lambda}\right), \tau_0\left(x-X\sqrt{\lambda}\right)}{\sqrt{\lambda}}\right)\right\},$$

and using the fact that $|\tau'(x_0)|<1$, this set is bounded under the condition (164).

Thus, immediately,

$$|u(x,t)-u_0(x,t)| = o\left(\sqrt{\lambda}\right). \tag{165}$$

A consequence of (165) is that, for λ sufficiently small, the solution u of (147) is negative on the set

$$T^-_{x_0, t_0+(\pi-\varepsilon)\sqrt{\lambda(1-m^2)}} \cap \{(x,t): t \ge \tau(x)\}. \tag{166}$$

This uses the fact that $u_t<0$ on a neighborhood of x_0, as was proved in Lemma 19.

Therefore, on the set (166), the solution of the penalized problem (122) is the solution of the linear problem (147), for λ small enough. We have thus, for (x,t) on the set (166):

$$\frac{\partial u_\lambda}{\partial t}(x',t') = \frac{\partial w}{\partial t}(x',t') - \frac{1}{2\lambda}\int_{(x',t-|x-x'|)\in S} \frac{\partial w}{\partial t}(x', t-|x-x'|)\,dx' - \frac{1}{2\lambda\sqrt{\lambda}} \int_{T^-_{x,t}} \mathcal{G}_0'\left(\frac{\sqrt{(t-t')^2-(x-x')^2}}{\sqrt{\lambda}}\right)(t-t')(w\cdot 1_S)(x,t)\,dx'\,dt.$$

Reasoning as for (165), we can prove under assumption (164) that

$$\left|\left(\frac{\partial u_\lambda}{\partial t} - \frac{\partial u_0}{\partial t}\right)(x,t)\right| = o(1),$$

or

$$\left|\frac{\partial u_\lambda}{\partial t}(x,t) - w_t(x_0,t_0)\cos\frac{t-t_0-m(x-x_0)}{\sqrt{\lambda(1-m^2)}}\right| = o(1),$$

and, in particular

$$\lim_{\lambda\to 0} \frac{\partial u_\lambda}{\partial t}\left(x_0, t_0+(\pi-\varepsilon)\sqrt{\lambda(1-m^2)}\right) = +w_t(x_0,t_0)\cos(\pi-\varepsilon) \tag{167}$$

$$\text{on } \{x_0 : \tau'(x_0)<1 \text{ and } w_t(x_0, \tau(x_0); Q_i)<0, i=1,\cdots,4\}.$$

Analogously,

$$\lim_{\lambda\to 0}\frac{\partial u_\lambda}{\partial x}\left(x_0,t_0+(\pi-\varepsilon)\sqrt{\lambda(1-m^2)}\right)=w_x(x_0,t_0)\cos(\pi-\varepsilon) \tag{168}$$

$$\text{on } \{x_0:\tau'(x_0)<1 \text{ and } w_t(x_0,\tau(x_0);Q_i)<0, i=1,\cdots,4\},$$

and (167) and (168) in turn imply:

$$\begin{aligned}&\lim_{\lambda\to 0}\frac{\partial u_\lambda}{\partial \xi}\left(x_0,t_0+(\pi-\varepsilon)\sqrt{\lambda(1-m^2)}\right)=w_\xi(x_0,t_0)\cos(\pi-\varepsilon),\\ &\lim_{\lambda\to 0}\frac{\partial u_\lambda}{\partial \eta}\left(x_0,t_0+(\pi-\varepsilon)\sqrt{\lambda(1-m^2)}\right)=w_\eta(x_0,t_0)\cos(\pi-\varepsilon).\end{aligned} \tag{169}$$

Therefore, the limit $\bar{u}$ of u satisfies:

$$\frac{\partial \bar{u}}{\partial t}(x,\tau(x))=-\frac{\partial w}{\partial t}(x,\tau(x)) \quad \text{a.e. on } \{x:|\tau'(x)|<1\}.$$

This proves that $\bar{u}$ is indeed the solution of (P_∞). □

6. Conclusion. There are still many open problems which can be conveniently listed at this point. The main one is to prove existence of an energy conserving solution when the obstacle is not assumed to be concave, as was the case in [5].

An obstruction to the proof of existence is that the lines of influence might cluster, and we do not know how to extend the solution after they have clustered.

But there is a more fundamental problem: the whole model relies on the assumption that the motion is transverse: how well is this assumption satisfied when the obstacle is not parallel to the rest position of the string? A better model might be needed; it should be at the same time realistic and tractable.

Another class of problems is the study of the qualitative properties of the system that we consider: periodicity, almost periodicity, for instance; for a first set of results in this direction, see [7].

REFERENCES

[1] L. Amerio, *Su un problema di vincoli unilaterali per l'equazione non omogenea della corda vibrante*, Pubbl. IACD, 190 (1976), pp. 3–11.

[2] A. Bamberger, personal communication, 1978.

[3] _____, Thèse d'état, Université P. et M. Curie, Paris, 1978.

[4] C. Citrini, *Discontinuous solutions of a nonlinear hyperbolic equation with unilateral constraints*, Manuscripta Math., 29 (1979), pp. 323–352.

[5] M. Schatzman, *A hyperbolic problem of second order with unilateral constraints: the vibrating string with a concave obstacle*, J. Math. Anal. Appl., 73 (1980), pp. 138–191.

[6] _____, *Un problème hyperbolique du 2ème ordre avec constraintes unilatérales: la corde vibrante avec obstacle ponctuel*, J. Differential Equations, 36 (1980), pp. 295–334.

[7] H. Cabannes and A. Haraux, to appear.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0013 $01.25/0

HOLOMORPHIC FUNCTIONS IN TUBES WHICH HAVE DISTRIBUTIONAL BOUNDARY VALUES AND WHICH ARE H^p FUNCTIONS*

RICHARD D. CARMICHAEL† AND STEPHEN P. RICHTERS†

Abstract. Let C be an open convex cone in $\mathbf{R}^n$, n-dimensional real space, such that $\overline{C}$ does not contain any entire straight line. We consider holomorphic functions in tubes $T^C=\mathbf{R}^n+iC\subset\mathbf{C}^n$, n-dimensional complex space, which are known to have $\mathcal{S}'$ distributional boundary values. We prove that if the $\mathcal{S}'$ boundary value of such a holomorphic function is an L^p function, $1\leq p\leq\infty$, then the holomorphic function is in the Hardy space $H^p(T^C)$, $1\leq p\leq\infty$, corresponding to the tube T^C and can be recovered as the Poisson integral of the distributional boundary value. A similar result is proved wih respect to the holomorphic functions in tubes T^C which are known to have boundary values in the distribution spaces $\mathcal{Z}'$, K'_r, $(\mathcal{S}^\alpha)'$, and $(W^\Omega)'$ and which do not necessarily have S' distributional boundary values. In all cases we prove converse theorems. Our basic results are motivated by a recent 1-dimensional theorem which is associated with calculations in theoretical physics. Our results extend this 1-dimensional theorem to a much more general setting and are also obtained with respect to two types of growth conditions on the holomorphic functions and with respect to several distribution topologies as noted above. In addition, as part of the analysis needed to prove our basic theorems we have obtained some new results concerning the Hardy $H^p(T^C)$-spaces and Poisson integrals corresponding to tubes T^C and have also obtained some new distributional boundary value results.

1. Introduction. Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. We consider functions which are holomorphic in the tube $T^c=\mathbf{R}^n+iC\subset\mathbf{C}^n$ and which satisfy a growth condition under which the functions are known to obtain distributional boundary values in $\mathcal{S}'$, the space of tempered distributions. We prove that if the $\mathcal{S}'$ boundary value of such a holomorphic function is an L^p function, $1\leq p\leq\infty$, then the holomorphic function is in the Hardy space $H^p(T^C)$ and can be recovered as the Poisson integral of the distributional boundary value. We then consider holomorphic functions in tubes T^C which satisfy a growth under which the functions are known to have distributional boundary values in the spaces $\mathcal{Z}'$, K'_r, $(\mathcal{S}^\alpha)'$ and $(W^\Omega)'$ but do not necessarily have boundary values in $\mathcal{S}'$; again we prove that if the distributional boundary value is in L^p, $1\leq p\leq\infty$, then the holomorphic function is in $H^p(T^C)$ and is recoverable by the Poisson integral of the boundary value. In all cases we prove converse results.

This paper has been motivated by [28, Thm. 2] which is a 1-dimensional result like that described in the preceding paragraph and which becomes a special case of our work here. In [28] Raina notes the importance of the H^p spaces in particle physics and gives several references in which the H^p spaces are used in applicable calculations. The distributional boundary value computation, with which [28, Thm. 2] is concerned with respect to the distribution space $(\mathcal{S}^1)'$, is important in quantum field theory where the vacuum expectation values are distributional boundary values in a relevant distribution topology of holomorphic functions defined in subsets of $\mathbf{C}^n$. In [28, Thm. 2] the existence of an $(\mathcal{S}^1)'$ boundary value follows from the growth [28, (3.2)] by the analysis of Constantinescu [13], which we have generalized to arbitrary tubes in [10]. Constantinescu constructs local fields, which are a category of fields larger than the strictly localizable ones and which also contain the tempered fields; he proves that the vacuum expectation values in local fields are distributional boundary values in $(\mathcal{S}^1)'$ of holomorphic functions in a tube domain in $\mathbf{C}^n$ corresponding to the forward light cone in

*Received by the editors March 24, 1981.
†Department of Mathematics, Wake Forest University, Winston-Salem, North Carolina 27109.

$\mathbf{R}^n$. (See [13] also for references and discussion of other field theories.) Other domains of holomorphicity in applications in the computation of distributional boundary values are the half planes in $\mathbf{C}^1$, the tube in $\mathbf{C}^n$ corresponding to the backward light cone, any of the 2^n generalized half planes in $\mathbf{C}^n$ corresponding to the quadrants in $\mathbf{R}^n$, and more generally tubes in $\mathbf{C}^n$ defined by open convex cones in $\mathbf{R}^n$ whose dual cones have nonempty interior. In the latter case, important and interesting distributional boundary value results for tempered distributions which have application in quantum field theory are obtained in [32, Chap. II, §II.2]; also see [29, §IX.3]. We also refer to [27] for important representation results of holomorphic functions in tubes by Fourier–Laplace transforms of tempered distributions.

Because of the importance in mathematics and in theoretical physics of the Hardy H^p spaces and of the distributional boundary value computation for holomorphic functions in tubes in $\mathbf{C}^n$ which satisfy various growth conditions, we prove in this paper the results indicated in the first paragraph of this section. In so doing we also prove some new results concerning the H^p functions and Poisson integrals corresponding to tubes T^C in $\mathbf{C}^n$ and obtain new distributional boundary value results for several spaces of distributions.

In §2 of this paper we state notation and definitions and obtain some technical results which will be needed. Cauchy and Poisson integrals and H^p spaces corresponding to tubes T^C in $\mathbf{C}^n$ are discussed in §3 where we obtain several needed results concerning these topics. We define the holomorphic functions in tubes T^C which generalize the functions considered in [28, Thm. 2] in §4 and state some new distributional boundary value results corresponding to these holomorphic functions which extend some previous results of ours. The growth of holomorphic functions in tubes which have $\mathcal{S}'$ boundary values and which we study in this paper is also defined in §4. Sections 5 and 6 are devoted to obtaining the basic results of this paper as described in the first paragraph of this section corresponding to the various generalized function topologies and corresponding to the two types of growths on the holomorphic functions which we consider here.

2. Notation, definitions, and technical results. The n-dimensional notation to be used in this paper is exactly that described in [3, p. 1042]. Note especially the n-dimensional differential operators D_t^α and D_z^α. If $z\in\mathbf{C}^n$, $|z|$ is as defined in [6, p. 844]. $\bar{0}$ will denote the n-tuple of zeros, $\bar{0}=(0,0,\cdots,0)$, throughout the paper.

The definitions of a cone C in $\mathbf{R}^n$, projection of a cone C, $\mathrm{pr}(C)$, compact subcone C' of a cone C, and the indicatrix function of a cone C, $u_C(t)$, are given in [3, p. 1042]. The dual cone C^* of a cone C is defined to be $C^*=\{t\in\mathbf{R}^n:\langle t,y\rangle\geq 0, \text{ for all } y\in C\}=\{t\in\mathbf{R}^n: u_C(t)\leq 0\}$. $O(C)$ shall denote the convex envelope of the cone C, and T^C denotes the tube $T^C=\mathbf{R}^n+iC$ in $\mathbf{C}^n$ defined by the cone C.

The L^1 Fourier and inverse Fourier transforms are defined in [3, p. 1042]. The limit in the mean Fourier and inverse Fourier transforms of functions in L^p, $1<p\leq 2$, and L^q, $(1/p)+(1/q)=1$, are in [21] and [2]; and we assume familiarity on the part of the reader with the properties, such as the Parseval equality and inequality, of these transforms in n dimensions. $\mathcal{F}[\phi(t);x]$ ($\mathcal{F}^{-1}[\phi(x);t]$) shall denote the Fourier (inverse Fourier) transform of a function ϕ in the relevant sense throughout the paper.

The function spaces $\mathcal{D},\mathcal{S},\mathcal{K}_p,\mathcal{S}_\alpha$, and W_M and their respective Fourier transform spaces $\mathcal{Z},\mathcal{S},K_p,\mathcal{S}^\alpha$, and W^Ω along with the respective properties and topologies can be found in [17], [30], [3], [18] or [10], and [19] or [26], respectively. The function spaces $\mathcal{E},\mathcal{D}_{L^p}$, and $\mathcal{B}$ can be found in [30]; also see [4] and [9]. The same references as above also discuss the corresponding distribution spaces $\mathcal{D}',\mathcal{S}',\mathcal{K}_p',\mathcal{S}_\alpha'$, and W_M' and their respective Fourier transform spaces $\mathcal{Z}',\mathcal{S}',K_p'$, $(\mathcal{S}^\alpha)'$, and $(W^\Omega)'$. We ask the reader to

note these references for the respective definitions of the distributional Fourier (inverse Fourier) transform relating $\mathcal{D}' \leftrightarrow \mathcal{L}'$, $\mathcal{S}' \leftrightarrow \mathcal{S}'$, $\mathcal{H}_p' \leftrightarrow K_p'$, $\mathcal{S}_\alpha' \leftrightarrow (\mathcal{S}^\alpha)'$, and $W_M' \leftrightarrow (W^\Omega)'$ in a linear, continuous, one-to-one, and onto manner in each case. In this paper $\mathcal{F}[V]$ ($\mathcal{F}^{-1}[U]$) will denote the Fourier (inverse Fourier) transform of a distribution or generalized function V (U). In our transforms we delete the factor $(2\pi)^n$ which is contained in [17, p. 190, (1)], for example, because of the way we have defined the Fourier transform of functions with 2π in the exponent of the exponential term.

All definitions and terminology concerning distributions, such as the support of a distribution, will be that of Schwartz [30]. The support of a function f and of a distribution V will be denoted by $\operatorname{supp}(f)$ and $\operatorname{supp}(V)$. All terminology from topological vector spaces and their dual spaces, such as bounded set in a topological vector space and strongly bounded set in a dual space, can be found in [14] and [16, Chap. 1].

Let C be an open connected cone in $\mathbf{R}^n$, and let C' be an arbitrary compact subcone of C. Let $f(z)$ be a function of $z=x+iy\in T^C$. Let U be a distribution or generalized function. By $f(x+iy)\to U$ in the weak topology of the distribution space as $y\to\bar{0}$, $y\in C$, we mean $\langle f(x+iy),\psi(x)\rangle\to\langle U,\psi\rangle$ as $y\to\bar{0}$, $y\in C'\subset C$, for every compact subcone C' of C and for each fixed element ψ in the corresponding test function space. By $f(x+iy)\to U$ in the strong topology of the distribution space as $y\to\bar{0}$, $y\in C$, we mean $\langle f(x+iy),\psi(x)\rangle\to\langle U,\psi\rangle$ as $y\to\bar{0}$, $y\in C'\subset C$, for every compact subcone C' of C where the convergence is uniform for ψ on arbitrary bounded sets in the corresponding test function space. U is then called the weak or strong, respectively, distributional boundary value of $f(z)$; this boundary value is defined on the distinguished boundary of the tube T^C, $\{z=x+iy : x\in\mathbf{R}^n,\ y=\bar{0}\}$, which is not necessarily the topological boundary of T^C.

In the remainder of this section we prove technical results which will be useful in this paper.

LEMMA 2.1. *Let C be an open connected cone in $\mathbf{R}^n$ and let $I_{C^*}(t)$ denote the characteristic function of the dual cone C^* of C. We have*

$$(I_{C^*}(t)\exp(2\pi i\langle z,t\rangle))\in L^p \quad \text{for all } p,\quad 1\le p\le\infty, \tag{2.1}$$

as a function of $t\in\mathbf{R}^n$ for arbitrary $z\in T^{O(C)}$.

Proof. See [8, Lemma 2]. (The word "volume" in [8, p. 577, line 11] should be "surface area".)

Let $\mu=(\mu_1,\mu_2,\cdots,\mu_n)$ be any of the 2^n n-tuples whose entries are 0 or 1. Let $z=x+iy\in T^{C_\mu}=\mathbf{R}^n+iC_\mu$ where $C_\mu=\{y\in\mathbf{R}^n:(-1)^{\mu_j}y_j>0,\ j=1,\cdots,n\}$ is any one of the 2^n quadrants in $\mathbf{R}^n$. Put

$$X(z)=\prod_{j=1}^{n}\left(1-i(-1)^{\mu_j}z_j\right)^{N+n+2},\qquad z\in T^{C_\mu}, \tag{2.2}$$

where $N\ge 0$ is fixed and n is the dimension.

LEMMA 2.2. *For the function $X(z)$, $z\in T^{C_\mu}$, defined in* (2.2) *we have $X(x+iy)\to X(x)$ in the weak and strong topologies of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C_\mu$. Further, if there exist elements $U\in\mathcal{S}'$ and $h(x)\in L^p$, $1\le p\le\infty$, such that $(X(x)U)=h(x)$ in $\mathcal{S}'$ then $U=(h(x)/X(x))$ in $\mathcal{S}'$.*

Proof. We desire a limit result as $y\to\bar{0}$, $y\in C_\mu$; thus without loss of generality we can assume $|y|\le M$, $y\in C_\mu$, for a fixed $M>0$. We have

$$\begin{aligned}|X(x+iy)-X(x)|&\le|X(x+iy)|+|X(x)|\\ &\le\prod_{j=1}^{n}\left((1+M)^2+x_j^2\right)^{(N+n+2)/2}+\prod_{j=1}^{n}\left(1+x_j^2\right)^{(N+n+2)/2}\end{aligned} \tag{2.3}$$

since $((-1)^{\mu_j}y_j)>0, j=1,\cdots,n$. The right side of (2.3) when multiplied by an element $\phi\in\mathcal{S}$ is an L^1 function of $x\in\mathbf{R}^n$ which is independent of $y\in C_\mu$ such that $|y|\le M$. Using this fact and the fact that $X(x+iy)\to X(x)$ pointwise as $y\to\bar{0}$, $y\in C_\mu$, the Lebesgue dominated convergence theorem yields

$$\lim_{\substack{y\to\bar{0}\\ y\in C_\mu}}\int_{\mathbf{R}^n}X(x+iy)\phi(x)dx=\int_{\mathbf{R}^n}X(x)\phi(x)dx,\qquad \phi\in\mathcal{S},$$

which proves that $X(x+iy)\to X(x)$ in the weak topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C_\mu$. But $\mathcal{S}$ is a Montel space ([36, p. 21] or [14, p. 510].) Hence by [14, p. 510, Cor. 8.4.9] the convergence of $X(x+iy)$ to $X(x)$ holds in the strong topology of $\mathcal{S}'$ also.

To obtain the second result in Lemma 2.2 first note that $(1/X(x))$ is a multiplier in $\mathcal{S}$ and hence in $\mathcal{S}'$; and $h\in L^p$, $1\le p\le\infty$, implies $h\in\mathcal{S}'$. By hypothesis $(X(x)U)=h(x)$ in $\mathcal{S}'$, thus for any $\phi\in\mathcal{S}$

$$\langle\frac{h(x)}{X(x)},\phi(x)\rangle=\langle h(x),\frac{\phi(x)}{X(x)}\rangle=\langle X(x)U,\frac{\phi(x)}{X(x)}\rangle=\langle U,\phi\rangle,$$

which proves that $U=(h(x)/X(x))$ in $\mathcal{S}'$ as desired. The proof of Lemma 2.2 is complete.

LEMMA 2.3. *Let $X(z)$ be the function defined in* (2.2). *For each C_μ we have*

$$\left|\frac{1}{X(z)}\right|\le 1,\qquad z\in T^{C_\mu}. \tag{2.4}$$

If $((-1)^{\mu_j}y_j)\ge\delta_j>0$ for $j=1,\cdots,n$, $y\in C_\mu$, then

$$|X(x+iy)|\ge\prod_{j=1}^{n}\left(K_j(\delta_j)\right)^{N+n+2}(1+|z_j|)^{N+n+2} \tag{2.5}$$

for some constants $K_j(\delta_j)$ depending on δ_j, $j=1,\cdots,n$.

Proof. $z\in T^{C_\mu}$ implies $((-1)^{\mu_j}y_j)>0, j=1,\cdots,n$. Thus for each $j=1,\cdots,n$,

$$\left|1-i(-1)^{\mu_j}z_j\right|=\left(\left(1+(-1)^{\mu_j}y_j\right)^2+x_j^2\right)^{1/2}\ge\left(1+x_j^2\right)^{1/2}\ge 1$$

from which (2.4) follows.

To prove (2.5) note that for $z\in T^{C_\mu}$, $((1+|z_j|)/|z_j|)\to 1$ as $|z_j|\to\infty$, $j=1,\cdots,n$. Applying the definition of limit with $\varepsilon=\frac{1}{2}$ for each $j=1,\cdots,n$ we obtain a number $M_j>0$ such that

$$1+|z_j|<\tfrac{3}{2}|z_j|\quad\text{if } |z_j|\ge M_j>0. \tag{2.6}$$

If $|z_j|\le M_j$, $z\in T^{C_\mu}$, then for $((-1)^{\mu_j}y_j)\ge\delta_j>0$ we have

$$\frac{1+|z_j|}{|z_j|}\le\frac{1+M_j}{\delta_j}. \tag{2.7}$$

Combining (2.6) and (2.7) we have

$$|z_j|\ge K_j(\delta_j)(1+|z_j|)\quad\text{if } \left((-1)^{\mu_j}y_j\right)\ge\delta_j>0,\quad j=1,\cdots,n,$$

where $K_j(\delta_j)=\min\{2/3,\delta_j/(1+M_j)\}$. Thus for $y_j=\mathrm{Im}(z_j)$ satisfying $((-1)^{\mu_j}y_j)\ge\delta_j>0$, $j=1,\cdots,n$, we have

$$\left|1-i(-1)^{\mu_j}z_j\right|=\left(\left(1+(-1)^{\mu_j}y_j\right)^2+x_j^2\right)^{1/2}>|z_j|\ge K_j(\delta_j)(1+|z_j|)$$

for each $j=1,\cdots,n$. (2.5) follows from this inequality and the definition of $X(z)$. This completes the proof of Lemma 2.3.

Let $b\geq 0$ be fixed. Let $g(r)\in\mathcal{E}$, $r\in\mathbf{R}^1$, such that $g(r)=1$ if $r\geq(-b)$, $g(r)=0$ if $r\leq(-b-\varepsilon)$, $\varepsilon>0$ and fixed, and $0\leq g(r)\leq 1$. Let C be an open connected cone and put

$$\lambda(t)=g(\langle y,t\rangle),\qquad t\in\mathbf{R}^n,\quad y\in O(C). \tag{2.8}$$

We have $\lambda(t)\in\mathcal{E}$, $t\in\mathbf{R}^n$, for each $y\in O(C)$.

We now present a discussion which will be useful in the proof of Theorem 5.1 below. Let $h(x)\in L^p$, $1\leq p\leq 2$. Then $(h(x)/X(x))\in L^p$ since $|1/X(x)|\leq 1$, $x\in\mathbf{R}^n$, for $X(x)$ defined by (2.2) with $y=\bar{0}$ there. For $1<p\leq 2$, $H(t)=\mathcal{F}^{-1}[h(x)/X(x);t]\in L^q$ exists in the L^q sense, $(1/p)+(1/q)=1$; and if $p=1$, $H(t)=\mathcal{F}^{-1}[h(x)/X(x);t]\in L^\infty$. Then $H(t)=\mathcal{F}^{-1}[h(x)/X(x)]$ as elements of $\mathcal{S}'$ also, $1\leq p\leq 2$. Now assume that there exists an element $U\in\mathcal{S}'$ such that $U=(h(x)/X(x))$ in $\mathcal{S}'$ and that there exists a continuous function $G(t)$ having support in C^*, the dual cone of some given open convex cone C, such that $U=\mathcal{F}[G(t)]$ in $\mathcal{S}'$. We then have

$$\begin{aligned}\mathcal{F}[G(t)]=U&=\frac{h(x)}{X(x)}=\mathcal{F}[H(t)],\\ G(t)=\mathcal{F}^{-1}[U]&=\mathcal{F}^{-1}\left[\frac{h(x)}{X(x)}\right]=H(t)\end{aligned} \tag{2.9}$$

as equalities in $\mathcal{S}'$. Since $\operatorname{supp}(G)\subseteq C^*$ then $\operatorname{supp}(H)\subseteq C^*$ almost everywhere as a function. Now take $\lambda(t)$ defined in (2.8) corresponding to the present open convex cone C and with $b=0$ in the definition of (2.8). We have $(\lambda(t)\exp(2\pi i\langle z,t\rangle))\in\mathcal{S}$ as a function of $t\in\mathbf{R}^n$ for each $z\in T^C$. Using this fact, the assumption that $U=(h(x)/X(x))$ in $\mathcal{S}'$, (2.9), the Fourier and inverse Fourier transforms on $\mathcal{S}'$, and assuming that the integral on the left of (2.10) below is well defined (for it is later in the paper where we need this calculation), we have for all $z\in T^C$ that

$$\begin{aligned}\int_{C^*}G(t)e^{2\pi i\langle z,t\rangle}\,dt&=\langle G(t),\lambda(t)e^{2\pi i\langle z,t\rangle}\rangle=\langle U,\mathcal{F}^{-1}[\lambda(t)e^{2\pi i\langle z,t\rangle};\eta]\rangle\\ &=\left\langle\frac{h(\eta)}{X(\eta)},\mathcal{F}^{-1}[\lambda(t)e^{2\pi i\langle z,t\rangle};\eta]\right\rangle\\ &=\langle H(t),\lambda(t)e^{2\pi i\langle z,t\rangle}\rangle=\int_{C^*}H(t)e^{2\pi i\langle z,t\rangle}\,dt,\end{aligned} \tag{2.10}$$

and the last integral is well defined by Lemma 2.1 and the facts that $H(t)\in L^\infty$ if $p=1$ and $H(t)\in L^q$, $(1/p)+(1/q)=1$, if $1<p\leq 2$. We need the discussion in this paragraph and (2.10) in the proof of Theorem 5.1 below.

3. Cauchy and Poisson integrals and H^p spaces. We shall prove the main results of this paper, which are contained in §§5 and 6, corresponding to open convex cones C in $\mathbf{R}^n$ such that $\overline{C}$ does not contain any entire straight line. Because of this and for certain technical reasons, we assume that C is such a cone throughout this section.

The Cauchy kernel corresponding to the tube $T^C=\mathbf{R}^n+iC$ is

$$K(z-t)=\int_{C^*}\exp(2\pi i\langle z-t,\eta\rangle)\,d\eta,\qquad t\in\mathbf{R}^n,\quad z\in T^C, \tag{3.1}$$

where C^* is the dual cone of C. The Poisson kernel corresponding to T^C is

$$Q(z;t)=\frac{K(z-t)\overline{K(z-t)}}{K(2iy)}=\frac{|K(z-t)|^2}{K(2iy)}, \qquad t\in\mathbf{R}^n, \quad z\in T^C. \tag{3.2}$$

Because of [36, Lemma 1, p. 222] we need the assumption on $\bar{C}$ stated in the first paragraph of this section in order for $Q(z;t)$ to be well defined. We have obtained properties of $K(z-t)$ and $Q(z;t)$ in [4], [8], and [9]. Korányi [22, Prop. 2] and Stein and Weiss [33, p. 105] have noted that the Poisson kernel $Q(z;t)$ is an approximate identity; see also [4, Lemma 6, p. 213]. In the following lemma we collect facts from these references which we need in this paper.

LEMMA 3.1. *$K(z-t)$ is a holomorphic function of $z\in T^C$ for fixed $t\in\mathbf{R}^n$. For $1\le p\le 2$ and fixed $z\in T^C$, $K(z-t)\in\dot{\mathcal{B}}\cap\mathcal{D}_{L^q}$ for all q, $(1/p)+(1/q)=1$, and $Q(z;t)\in\dot{\mathcal{B}}\cap\mathcal{D}_{L^q}$ for all q, $1\le q\le\infty$, as functions of $t\in\mathbf{R}^n$. Further, $Q(z;t)$ satisfies the following approximate identity properties*:

$$Q(z;t)\ge 0, \qquad t\in\mathbf{R}^n, \quad z\in T^C, \tag{3.3}$$

$$\int_{\mathbf{R}^n} Q(z;t)\,dt=1, \qquad z\in T^C, \tag{3.4}$$

if $\delta>0$,

$$\lim_{\substack{z\to t_0\\ z\in T^C}}\int_{|t-t_0|>\delta} Q(z;t)\,dt=0 \tag{3.5}$$

uniformly for all $t_0\in\mathbf{R}^n$.

We obtain some additional facts and calculations which we need concerning the Cauchy and Poisson kernel functions in the next three lemmas.

LEMMA 3.2. *Let $w=u+iv\in T^C$ be fixed. Then $K(z+w)$ is holomorphic in $z\in T^C$ and*

$$|K(z+w)|\le M_v<\infty, \qquad z\in T^C, \tag{3.6}$$

where M_v is a constant which depends only on $v=\mathrm{Im}(w)$. Further, we have that $K(x+iy+w)\to K(x+w)$ in the weak and strong topology of $\mathcal{S}'$ as $y=\mathrm{Im}(z)\to\bar{0}$, $y\in C$, for each $w\in T^C$.

Proof. For $w\in T^C$ fixed, $K(z+w)$ is holomorphic in $z\in T^C$ as in Lemma 3.1. Applying [8, Lemma 1] corresponding to $v=\mathrm{Im}(w)\in C$ we obtain $\delta=\delta_v>0$ depending on v such that $\langle v,\eta\rangle\ge\delta|v||\eta|$ for all $\eta\in C^*$, and by the same result $\langle y,\eta\rangle>0$, $y\in C$, $\eta\in C^*$. Using the definition of $K(z+w)$ from (3.1), [31, Thm. 32, p. 39], and integration by parts $(n-1)$ times we obtain for $z\in T^C$ that

$$\begin{aligned}|K(z+w)|&=\left|\int_{C^*}\exp(2\pi i\langle z+w,\eta\rangle)\,d\eta\right|\le\int_{C^*}e^{-2\pi\langle y,\eta\rangle}e^{-2\pi\langle v,\eta\rangle}\,d\eta\\ &\le\int_{C^*}e^{-2\pi\langle v,\eta\rangle}\,d\eta\le\int_{C^*}e^{-2\pi\delta|v||\eta|}\,d\eta\le\int_{\mathbf{R}^n}e^{-2\pi\delta|v||\eta|}\,d\eta\\ &=\Omega_n\int_0^\infty r^{n-1}e^{-2\pi\delta|v|r}\,dr=\Omega_n(n-1)!(2\pi\delta|v|)^{-n}\end{aligned} \tag{3.7}$$

where Ω_n is the surface area of the unit sphere in $\mathbf{R}^n$; and (3.6) is obtained with $M_v=\Omega_n(n-1)!(2\pi\delta|v|)^{-n}$.

For $w \in T^C$ and $z \in T^C$, the analysis of (3.7) yields

$$\begin{aligned}(3.8)\qquad &|I_{C^*}(\eta)(\exp(2\pi i\langle z+w,\eta\rangle)-\exp(2\pi i\langle x+w,\eta\rangle))| \\ &\qquad \le I_{C^*}(\eta)(e^{-2\pi\langle y,\eta\rangle}e^{-2\pi\langle v,\eta\rangle}+e^{-2\pi\langle v,\eta\rangle}) \le 2I_{C^*}(\eta)e^{-2\pi\langle v,\eta\rangle}\end{aligned}$$

where $I_{C^*}(\eta)$ is the characteristic function of C^*; and the right side of (3.8) is an L^1 function of $\eta \in \mathbf{R}^n$ by the analysis of (3.7) which is independent of $y = \mathrm{Im}(z) \in C$. Since $(I_{C^*}(\eta)(\exp(2\pi i\langle z+w,\eta\rangle)-\exp(2\pi i\langle x+w,\eta\rangle))) \to 0$ pointwise in $\eta \in \mathbf{R}^n$ as $y = \mathrm{Im}(z) \to \bar{0}$, $y \in C$, for each $x \in \mathbf{R}^n$, the Lebesgue dominated convergence theorem yields that $K(z+w) \to K(x+w)$ pointwise in $x \in \mathbf{R}^n$ as $y \to \bar{0}$, $y \in C$, for any fixed $w \in T^C$. Now let $\phi \in \mathcal{S}$. We similarly obtain as in (3.8) and (3.7) that

$$(3.9)\qquad |(K(z+w)-K(x+w))\phi(x)| \le 2M_v|\phi(x)|$$

where M_v is the constant in (3.6); and the right side of (3.9) is an L^1 function of $x \in \mathbf{R}^n$ which is independent of $y \in C$. Another application of the Lebesgue dominated convergence theorem now yields

$$\lim_{\substack{y \to \bar{0} \\ y \in C}} \int_{\mathbf{R}^n} (K(z+w)-K(x+w))\phi(x)\,dx = 0, \qquad \phi \in \mathcal{S},$$

for any $w \in T^C$. This proves the desired weak convergence of $K(z+w)$ to $K(x+w)$ in $\mathcal{S}'$. The strong $\mathcal{S}'$ convergence follows from this as in the proof of Lemma 2.2. The proof of Lemma 3.2 is complete.

LEMMA 3.3. *Let $g(t) \in L^p$, $1 \le p \le 2$, and let $G(\eta) = \mathcal{F}^{-1}[g(t);\eta]$ in the function sense, which exists. Assume that $(G(\eta)\exp(2\pi i\langle z,\eta\rangle)) \in L^1$ as a function of $\eta \in \mathbf{R}^n$ for $z \in T^C$ and that* $\mathrm{supp}(G) \subseteq C^*$ *almost everywhere. Then*

$$(3.10)\qquad \int_{C^*} G(\eta)e^{2\pi i\langle z,\eta\rangle}\,d\eta = \int_{\mathbf{R}^n} g(t)K(z-t)\,dt, \qquad z \in T^C.$$

Proof. The integral on the right of (3.10) is well defined for $1 \le p \le 2$ because of the properties of $K(z-t)$ noted in Lemma 3.1. For $1 < p \le 2$, (3.10) is obtained by using Lemma 2.1, Lemma 3.1, the definition of the inverse Fourier transform in L^q, $(1/p)+(1/q)=1$, Fubini's theorem, and exactly the same calculation as in [33, p. 104, lines 3–6], which holds equally well for $1 < p \le 2$. If $p = 1$, the fact that $G(\eta) = \mathcal{F}^{-1}[g(t);\eta]$ is the L^1 transform along with Lemmas 2.1 and 3.1 and a direct application of Fubini's theorem yields (3.10). We ask the reader to verify the details if desirable.

LEMMA 3.4. *Let z_0 be an arbitrary but fixed point in T^C. Let $1 \le p \le \infty$. There exists a closed neighborhood $N(z_0,\delta) = \{z : |z-z_0| \le \delta,\ \delta > 0\}$ of z_0 which is contained in T^C and a constant $B(z_0)$ depending only on z_0 such that*

$$(3.11)\qquad \|Q(z;t)\|_{L^p} \le B(z_0) < \infty, \qquad z \in N(z_0,\delta),$$

where the L^p norm is with respect to $t \in \mathbf{R}^n$.

Proof. From the definitions of $K(z-t)$ and $Q(z;t)$ in (3.1) and (3.2), respectively, and from (3.3) we have

$$(3.12)\qquad 0 \le Q(z;t) = \frac{|K(z-t)|^2}{K(2iy)} \le \frac{(K(iy))^2}{K(2iy)}, \qquad z = x+iy \in T^C, \quad t \in \mathbf{R}^n,$$

and by the analysis of Lemma 3.2, $0 < K(2iy) < \infty$, $y \in C$. Let $z_0 = x_0 + iy_0$ be an arbitrary but fixed point in T^C. Since C is open there exists a closed neighborhood of

z_0, $N(z_0,\delta)\subset T^C$, such that $\{y:|y-y_0|\leq\delta\}\subset C$. Now $K(iy)$ and $1/K(2iy)$ are continuous functions of y at each point of C. Thus $K(iy)$ and $1/K(2iy)$ are bounded on $\{y:|y-y_0|\leq\delta\}\subset C$ by constants depending only on y_0; hence from (3.12) we have

$$0\leq Q(z;t)\leq B'(z_0),\qquad z=x+iy\in N(z_0,\delta),\quad t\in\mathbf{R}^n, \tag{3.13}$$

where $B'(z_0)$ is a constant which actually depends only on $y_0=\mathrm{Im}(z_0)$. (3.13) proves (3.11) in the case $p=\infty$. Now let $1\leq p<\infty$. Again using the continuity of $K(iy)$ and $1/K(2iy)$ on C we have from (3.13) and (3.4) that for $z\in N(z_0,\delta)$

$$\int_{\mathbf{R}^n}|Q(z;t)|^p\,dt\leq\left(B'(z_0)\right)^{p-1}\int_{\mathbf{R}^n}Q(z;t)\,dt=\left(B'(z_0)\right)^{p-1}$$

and (3.11) follows with $B(z_0)=(B'(z_0))^{(p-1)/p}$. The proof is complete.

A function $f(z)$ which is holomorphic in the tube $T^C=\mathbf{R}^n+iC$ belongs to the Hardy class $H^p(T^C)$, $0<p<\infty$, ([33, pp. 90–91], [22, p. 276]) if there exists a constant $A<\infty$ which is independent of $y\in C$ such that

$$\int_{\mathbf{R}^n}|f(x+iy)|^p\,dx\leq A\quad\text{for all } y\in C. \tag{3.14}$$

The Hardy class $H^\infty(T^C)$ is the space of all bounded holomorphic functions in T^C.

The following lemma is a converse to [22, Prop. 4] and is already known in the special case that T^C is a half plane in $\mathbf{C}^1$.

LEMMA 3.5. *Let $f(z)$ be holomorphic in T^C and have the Poisson integral representation*

$$f(z)=\int_{\mathbf{R}^n}h(t)Q(z;t)\,dt,\qquad z\in T^C, \tag{3.15}$$

for some $h\in L^p$, $1\leq p\leq\infty$. Then $f(z)\in H^p(T^C)$, $1\leq p\leq\infty$; and $f(x+iy)\to h(x)$ as $y\to\bar{0}$, $y\in C$, in L^p if $1\leq p<\infty$ and in the weak-star topology of L^∞ if $p=\infty$.

Proof. The integral in (3.15) is well defined for $1\leq p\leq\infty$ because of Lemma 3.1. First let $h\in L^\infty$. Using (3.4) we have

$$|f(x+iy)|\leq A\int_{\mathbf{R}^n}Q(z;t)\,dt=A,\qquad z\in T^C,$$

where A is a bound on $h\in L^\infty$ almost everywhere; hence $f(z)\in H^\infty(T^C)$. If $1\leq p<\infty$ we use Jensen's inequality [15, 2.4.19, p. 91], the approximate identity properties of $Q(z;t)$ given in Lemma 3.1, and Fubini's theorem to obtain

$$\begin{aligned}\int_{\mathbf{R}^n}|f(x+iy)|^p\,dx&\leq\int_{\mathbf{R}^n}\int_{\mathbf{R}^n}|h(t)|^pQ(z;t)\,dt\,dx\\&=\int_{\mathbf{R}^n}|h(t)|^p\int_{\mathbf{R}^n}Q(z;t)\,dx\,dt=\int_{\mathbf{R}^n}|h(t)|^p\,dt\end{aligned} \tag{3.16}$$

for all $z=x+iy\in T^C$. (Note that the integral of $Q(z;t)$, $z=x+iy\in T^C$, $t\in\mathbf{R}^n$, with respect to $x\in\mathbf{R}^n$ in (3.16) is 1. This fact follows by a change of variable and [33, (ii), p. 105] or [4, (50), p. 213] and hence also follows from (3.4).) (3.16) is the desired growth (3.14), and $f(z)\in H^p(T^C)$, $1\leq p<\infty$. The desired convergence of $f(x+iy)\to h(x)$ as $y\to\bar{0}$, $y\in C$, in L^p if $1\leq p<\infty$ and in the weak-star topology of L^∞ if $p=\infty$ follows from (3.15) and [22, Prop. 3(c) and 3(d), p. 280]; for the weak-star topology of L^∞ the reader is referred to [20, p. 8]. The proof is complete.

The following growth result for $H^p(T^C)$ functions, $1\leq p<\infty$, is needed in this paper and has been proved in [11]; we do not repeat the argument here. Of course we

have not included $p=\infty$ in the result since by definition the $H^\infty(T^C)$ functions are the bounded holomorphic functions in T^C. Because of the previously known growth for $H^p(T^C)$ functions, $0<p\leq\infty$, in the special cases that $C=(0,\infty)$ or $(-\infty,0)$ in 1 dimension or $C=C_\mu$, any of the 2^n quadrants in n dimensions, as noted in (3.20) below, our growth (3.17) below is of interest for dimensions $n\geq 2$ and for cones that are not the quadrants.

THEOREM 3.1. *Let $f(z)\in H^p(T^C)$, $1\leq p<\infty$, where C is an open convex cone in $\mathbf{R}^n$ such that $\overline{C}$ does not contain any entire straight line. For any compact subcone C' of C there exists a constant $M(C')$ depending on C' such that*

$$|f(x+iy)|\leq M(C')|y|^{-n/p}, \qquad z=x+iy\in T^{C'}. \tag{3.17}$$

The following result provides the representation of H^p functions, $0<p\leq\infty$, in terms of the Fourier–Laplace integral. For $0<p\leq 1$ and $2<p\leq\infty$ the representations are new. The result is of independent interest; with respect to this paper we need case III of the result, which is already known, and the proof of case IV. Thus we give a detailed proof only of case IV.

THEOREM 3.2. *Let $f(z)\in H^p(T^C)$, $0<p\leq\infty$, where C is an open convex cone in $\mathbf{R}^n$ such that $\overline{C}$ does not contain any entire straight line.*

I. *If $0<p<1$ and $C=C_\mu$ is any of the 2^n quadrants, there exists $V\in\mathcal{S}'$ with* $\operatorname{supp}(V)\subseteq C_\mu^*=\overline{C}_\mu$ *such that*

$$f(z)=\langle V, e^{2\pi i\langle z,t\rangle}\rangle, \qquad z\in T^C. \tag{3.18}$$

II. *If $p=1$ there exists $V\in\mathcal{S}'$ with* $\operatorname{supp}(V)\subseteq C^*$ *such that* (3.18) *holds.*

III. *If $1<p\leq 2$ there exists $g\in L^q$, $(1/p)+(1/q)=1$, with* $\operatorname{supp}(g)\subseteq C^*$ *almost everywhere such that*

$$f(z)=\int_{\mathbf{R}^n} g(t)e^{2\pi i\langle z,t\rangle}\,dt, \qquad z\in T^C. \tag{3.19}$$

IV. *If $2<p\leq\infty$ and C is contained in or is any of the 2^n quadrants C_μ, there exists $V\in\mathcal{D}'_{L^2}$ with* $\operatorname{supp}(V)\subseteq C^*$ *such that* (3.18) *holds.*

V. *If $2<p\leq\infty$ there exists $V\in\mathcal{S}'$ with* $\operatorname{supp}(V)\subseteq C^*$ *such that* (3.18) *holds.*

From analysis of Madych [23], if $f(z)\in H^p(T^{C_\mu})$, $0<p\leq\infty$, for any quadrant C_μ, there exists a constant M, depending only on f, such that

$$|f(x+iy)|\leq M\left(\prod_{j=1}^n |y_j|\right)^{-1/p}, \qquad z=x+iy\in T^{C_\mu}. \tag{3.20}$$

Using (3.20), case I is proved by analysis from [5] and also can be obtained as a corollary to Theorem 4.4 below. Cases II and V follow from Theorems 3.1 and 4.4 of this paper. Case III is a special case of [9, Cor. 4.1 and 4.2]. In cases I, II, and V we also use the fact that $f(z)$ obtains an $\mathcal{S}'$ boundary value, a fact which follows from analysis contained in [36], [35], and [22] and applied to the various cases where relevant.

We have included case IV in Theorem 3.2 for two reasons. First the conclusion of the existence of an element $V\in\mathcal{D}'_{L^2}$ in case IV is somewhat more precise than can be obtained in case V for the more general cone since $\mathcal{D}'_{L^2}\subset\mathcal{S}'$. Secondly, as we have previously noted, we desire to display the proof of case IV for later reference in this paper, and we give this proof now.

Proof of Theorem 3.2 IV. For any cone C as hypothesized in case IV put

$$F(z)=\frac{f(z)}{Y(z)}, \qquad z\in T^C, \tag{3.21}$$

where

$$Y(z)=\prod_{j=1}^{n}\left(1-i(-1)^{\mu_j}z_j\right)^2, \qquad z\in T^C. \tag{3.22}$$

By analysis as in the proof of Lemma 2.3

$$\left|\frac{1}{Y(z)}\right|\leq\prod_{j=1}^{n}\left(1+x_j^2\right)^{-1}\leq 1, \qquad z=x+iy\in T^C; \tag{3.23}$$

hence $1/Y(x+iy)\in L^q$ for all q, $1\leq q\leq\infty$, as a function of $x=\mathrm{Re}(z)\in\mathbf{R}^n$ for $y\in C$ arbitrary. We are considering $2<p\leq\infty$ here. If $p=\infty$ we have from (3.23) and the hypothesis that $f(z)\in H^\infty(T^C)$ that

$$\int_{\mathbf{R}^n}|F(x+iy)|^2dx\leq A\int_{\mathbf{R}^n}\prod_{j=1}^{n}\left(1+x_j^2\right)^{-2}dx<\infty, \qquad y\in C, \tag{3.24}$$

where A is the bound on $f(z)\in H^\infty(T^C)$ which is independent of $y\in C$. Since $F(z)$ in (3.21) is holomorphic in T^C, (3.24) proves that $F(z)\in H^2(T^C)$ in the case $p=\infty$. For $2<p<\infty$ we use Hölders inequality and (3.23) to obtain for all $y\in C$ that

$$\begin{aligned}\int_{\mathbf{R}^n}|F(x+iy)|^2dx&\leq\|\,|f(x+iy)|^2\,\|_{L^{p/2}}\left\|\left|\frac{1}{Y(x+iy)}\right|^2\right\|_{L^{p/(p-2)}}\\&\leq A^{2/p}\left(\int_{\mathbf{R}^n}\left(\prod_{j=1}^{n}\left(1+x_j^2\right)^{-2p/(p-2)}\right)dx\right)^{(p-2)/p}\end{aligned} \tag{3.25}$$

where A is the constant in (3.14) corresponding to $f(z)\in H^p(T^C)$, $2<p<\infty$, here. The right side of (3.25) is a finite constant which is independent of $y\in C$. Thus again $F(z)\in H^2(T^C)$ for the case $2<p<\infty$. Thus for $2<p\leq\infty$ we apply case III of Theorem 3.2 or [33, Thm. 3.1, p. 101] and obtain a function $g(t)\in L^2$ with $\mathrm{supp}(g)\subseteq C^*$ almost everywhere such that

$$F(z)=\int_{\mathbf{R}^n}g(t)e^{2\pi i\langle z,t\rangle}dt, \qquad z\in T^C. \tag{3.26}$$

Now put

$$V=\left(\prod_{j=1}^{n}\left(1-i(-1)^{\mu_j}D_j\right)^2\right)(g(t))$$

where D_j is the differential operator with respect to t_j, $j=1,\cdots,n$, as noted in §2 above. (See [3, p. 1042].) We have $V\in\mathcal{D}'_{L^2}$ [30, Thm. XXV, p. 201], and $\mathrm{supp}(V)=\mathrm{supp}(g)\subseteq C^*$ as distributions since C^* is a regular set [30, pp. 98–99]. Taking $b=0$ in the definition of $\lambda(t)$ in (2.8) corresponding to our cone C here, we have $(\lambda(t)\exp(2\pi i\langle z,t\rangle))\in\mathcal{S}$ for $z\in T^C$. Noting that $\mathcal{D}'_{L^2}\subset\mathcal{S}'$ and recalling that $\mathrm{supp}(V)=\mathrm{supp}(g)\subseteq C^*$ we use distributional differention and obtain

$$\begin{aligned}\langle V,e^{2\pi i\langle z,t\rangle}\rangle&=\langle V,\lambda(t)e^{2\pi i\langle z,t\rangle}\rangle\\&=\prod_{j=1}^{n}\left(1-i(-1)^{\mu_j}z_j\right)^2\int_{\mathbf{R}^n}g(t)e^{2\pi i\langle z,t\rangle}dt, \qquad z\in T^C.\end{aligned} \tag{3.27}$$

Combining (3.21), (3.22), (3.26), and (3.27) we have (3.18). The proof of Theorem 3.2 IV is complete.

The final lemma in this section is a technical result which we need in §5. Before stating the lemma we note some further facts concerning the Poisson kernel function. Of course the integral in (3.4) is a continuous function of $z \in T^C$ since the integral is constant for all $z \in T^C$. For any fixed $M>0$ the integral

$$\int_{|t|\le M} Q(z;t)\,dt = \frac{1}{K(2iy)} \int_{|t|\le M} |K(z-t)|^2\,dt \tag{3.28}$$

is also a continuous function of $z=x+iy \in T^C$; hence so is

$$\int_{|t|>M} Q(z;t)\,dt = \int_{\mathbf{R}^n} Q(z;t)\,dt - \int_{|t|\le M} Q(z;t)\,dt. \tag{3.29}$$

We of course have

$$\lim_{M\to\infty} \int_{|t|\le M} Q(z;t)\,dt = \int_{\mathbf{R}^n} Q(z;t)\,dt = 1, \qquad z \in T^C. \tag{3.30}$$

LEMMA 3.6. *Let $h(t) \in L^\infty$. Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. Put*

$$X_\varepsilon(t) = \prod_{j=1}^{n} \left(1 - i\varepsilon(-1)^{\mu_j} t_j\right)^{N+n+2}, \qquad \varepsilon>0, \quad t \in \mathbf{R}^n, \tag{3.31}$$

where $N \ge 0$ is a fixed real number, n is the dimension, and $\mu=(\mu_1,\cdots,\mu_n)$ is any of the 2^n n-tuples whose entries are 0 or 1 that defines the quadrant C_μ. As $\varepsilon \to 0+$

$$H_\varepsilon(z) = \int_{\mathbf{R}^n} \left(\frac{h(t)}{X_\varepsilon(t)}\right) Q(z;t)\,dt \to H(z) = \int_{\mathbf{R}^n} h(t) Q(z;t)\,dt$$

uniformly in z on compact subsets of T^C.

Proof. Since $|1 - i\varepsilon(-1)^{\mu_j} t_j| \ge 1$, $j=1,\cdots,n$, for any $\varepsilon>0$ and any $t=(t_1,\cdots,t_n) \in \mathbf{R}^n$, then $|1/X_\varepsilon(t)| \le 1$ for all $\varepsilon>0$ and all $t \in \mathbf{R}^n$. Thus for all $\varepsilon>0$ and almost all $t \in \mathbf{R}^n$

$$\left|h(t) - \frac{h(t)}{X_\varepsilon(t)}\right| \le 2\|h\|_{L^\infty} = B. \tag{3.32}$$

Now let S denote any compact subset of T^C and let $\delta>0$ be arbitrary. Let z_0 be an arbitrary point in S. Because of (3.30), given $\delta>0$ there is a positive real number $M(\delta, z_0)$ depending on δ and on z_0 such that

$$\int_{|t|>M(\delta,z_0)} Q(z_0;t)\,dt < \frac{\delta}{4B} \tag{3.33}$$

where B is the number in (3.32). Using the continuity of integrals of the form on the left of (3.29) for any fixed $M>0$, given $\delta/4B$ there exists $\delta'(\delta, z_0)>0$ depending on δ and on z_0 such that

$$\left|\int_{|t|>M(\delta,z_0)} Q(z;t)\,dt - \int_{|t|>M(\delta,z_0)} Q(z_0;t)\,dt\right| < \frac{\delta}{4B}, \qquad |z-z_0|<\delta'(\delta,z_0). \tag{3.34}$$

Combining (3.33) and (3.34) we have

$$\int_{|t|>M(\delta,z_0)} Q(z;t)\,dt < \frac{\delta}{2B}, \qquad |z-z_0|<\delta'(\delta,z_0). \tag{3.35}$$

As z_0 varies over S, the balls $O(z_0,\delta'(\delta,z_0))=\{z:|z-z_0|<\delta'(\delta,z_0)\}$ cover S which is compact. Hence there is a finite number of these balls which cover S and which we enumerate as $O(z_j,\delta'(\delta,z_j))$, $j=1,\cdots,m$, with the z_j being the centers of the balls. Corresponding to each of these $z_j, j=1,\cdots,m$, there is the number $M(\delta,z_j), j=1,\cdots,m$, from (3.33) for which (3.35) holds. We now put

$$M(\delta,S)=\max\{M(\delta,z_1),M(\delta,z_2),\cdots,M(\delta,z_m)\}. \tag{3.36}$$

If z is any point in $S\subset T^C$, then z is in at least one of the $O(z_j,\delta'(\delta,z_j))$, $j=1,\cdots,m$; using this fact and (3.3) we then obtain from (3.35) and (3.36) that

$$\int_{|t|>M(\delta,S)}Q(z;t)\,dt\leq\int_{|t|>M(\delta,z_j)}Q(z;t)\,dt<\frac{\delta}{2B},\qquad z\in S\subset T^C, \tag{3.37}$$

where the $M(\delta,z_j)$ in (3.37) is chosen corresponding to $z\in S$ being in the ball $O(z_j,\delta'(\delta,z_j))$ for the appropriate $j=1,\cdots,m$. In (3.37) $M(\delta,S)>0$ depends only on $\delta>0$ and on the compact set $S\subset T^C$ and does not depend on $z\in S$. Recalling (3.3), (3.4), and (3.32) and using (3.37), we have for all $z\in S$ that

$$\begin{aligned}|H(z)-H_\varepsilon(z)|&\leq\int_{|t|\leq M(\delta,S)}\left|h(t)-\frac{h(t)}{X_\varepsilon(t)}\right|Q(z;t)\,dt\\&\quad+\int_{|t|>M(\delta,S)}\left|h(t)-\frac{h(t)}{X_\varepsilon(t)}\right|Q(z;t)\,dt\\&\leq\|h\|_{L^\infty}\sup_{|t|\leq M(\delta,S)}\left|1-\frac{1}{X_\varepsilon(t)}\right|\int_{|t|\leq M(\delta,S)}Q(z;t)\,dt\\&\quad+B\int_{|t|>M(\delta,S)}Q(z;t)\,dt\\&<\|h\|_{L^\infty}\sup_{|t|\leq M(\delta,S)}\left|1-\frac{1}{X_\varepsilon(t)}\right|+\frac{\delta}{2}\end{aligned} \tag{3.38}$$

for arbitrary $\delta>0$. Since $(1-(1/X_\varepsilon(t)))\to 0$ uniformly for t on compact subsets of $\mathbf{R}^n$ as $\varepsilon\to 0+$, the estimate (3.38) yields the desired result since S was any compact subset of T^C. The proof of Lemma 3.6 is complete.

4. Generalizations of previous results. In this section we state generalizations of some results concerning and related to distributional boundary values which we have obtained in previous papers. We do this because we desire to have these generalizations noted and because we need them to prove the main results of this paper. The proofs of these generalizations are obtained in the same way as the proofs which are already given for the special cases with only slight modifications. Thus we simply state our generalizations here and note that we have verified the modifications in the proofs of the special cases needed to prove the generalizations. The special case of each of the stated results in this section is noted in parenthesis next to the theorem number. We do have one result in this section which is new, and we indicate a proof of this.

Throughout this section $N(\bar{0},m)$ denotes the closed ball about $\bar{0}$ with radius $m>0$.

THEOREM 4.1. ([10, Lemma 10, p. 398]) *Let C be an open connected cone and let C' be an arbitrary compact subcone of C. Let $m>0$ be arbitrary but fixed. Let $g(t)$, $t\in\mathbf{R}^n$, be a continuous function with support in C^* which satisfies*

$$|g(t)|\leq M(C',m)\exp(2\pi(\langle\omega,t\rangle+\sigma|\omega|)),\qquad t\in\mathbf{R}^n, \tag{4.1}$$

for all $\sigma>0$, *where* $M(C',m)$ *is a constant which depends on* C' *and on* $m>0$ *and* (4.1) *is independent of* $\omega\in(C'\setminus(C'\cap N(\bar{0},m)))$ (*that is*, (4.1) *holds for all* $\omega\in(C'\setminus(C'\cap N(\bar{0},m)))$). *Let* y *be an arbitrary but fixed point of* C. *Then* $(\exp(-2\pi\langle y,t\rangle)g(t))\in L^p$ *for all* p, $1\leq p<\infty$, *as a function of* $t\in\mathbf{R}^n$.

We now define a class of holomorphic functions in tubes. Let C be an open connected cone and let $A\geq 0$ be a real number. For any real number $m>0$ and any compact subcone C' of C put $T(C';m)=\mathbf{R}^n+i(C'\setminus(C'\cap N(\bar{0},m)))$. We shall say that a function $f(z)$ belongs to the class $H(A;C)$ if $f(z)$ is holomorphic in the tube $T^C=\mathbf{R}^n+iC$ and if for every compact subcone C' of C and every $m>0$, there exists a constant $M(C',m)$ depending on C' and on $m>0$ such that

$$(4.2)\quad |f(x+iy)|\leq M(C',m)(1+|z|)^N\exp(2\pi(A+\sigma)|y|),\qquad z=x+iy\in T(C';m),$$

for all $\sigma>0$, where N is a nonnegative real number which does not depend on C' or on $m>0$.

The growth (4.2) is more general than [10, (23), p. 398] in that the constant $M(C',m)$ in (4.2) depends on C' and on $m>0$ instead of just on C'. Further, any function which is holomorphic in $T(C';m)$ for any compact subcone C' of C and any $m>0$ is also holomorphic in the whole of T^C; for if $z=x+iy\in T^C$, there exists a compact subcone C' of C and an $m>0$ such that $z\in T(C';m)$ since C is open. Thus if $f(z)$ is holomorphic in $T(C';m)$ for all $C'\subset C$ and all $m>0$, then $f(z)$ is holomorphic in T^C.

(4.2) is also more general than [6, (3), p. 845] and [7, (12), p. 772] because of the constant $M(C',m)$ and also because (4.2) holds in $T(C';m)$, $m>0$, and not necessarily in the whole of $T^{C'}$, $C'\subset C$. The functions $H(A;C)$ defined by the growth (4.2) are the correct functions to extend and generalize the functions considered by Raina [28, Thm. 2].

We should have defined the holomorphic functions considered in [6, §III], [7, §4], and [10, §4] to be the $H(A;C)$ functions in the first place because these are the more general and more natural functions with which to obtain the results there. We have already generalized the boundary value results of [7, §§4 and 5] to the $H(A;C)$ functions in [3, §8]. (See [3, §8, ¶1, p. 1062] for relevant discussion. The growth (4.2) is slightly more general than the growth [3, (6.4), p. 1053] which defines the functions $F_1(A;C)$ considered in [3, §8] because of the arbitrary $\sigma>0$ in (4.2). Nevertheless, the same proofs of [3, §8] yield the results of that section for the $H(A;C)$ functions also.) We now generalize results of [6] and [10] to the $H(A;C)$ functions and complete the correction of the lack of insight on our part now.

THEOREM 4.2. ([10, Lemma 11, pp. 399–400]) *Let* C *be an open connected cone. Let* $V=D^\gamma(g(t))$, *the distributional derivative of* $g(t)$ *of order* γ *with* γ *being an* n*-tuple of nonnegative integers, where* $g(t)$ *is a continuous function on* $\mathbf{R}^n$ *which satisfies* (4.1). *Let* $\operatorname{supp}(V)\subseteq C^*$. *Then* $f(z)=\langle V,\exp(2\pi i\langle z,t\rangle)\rangle$ *is an element of* $H(0;C)$.

THEOREM 4.3. ([6, Thm. 1, p. 846]) *Let* $f(z)\in H(A;C)$ *where* $A\geq 0$ *and* C *is an open connected cone. There exists a unique element* $U\in\mathcal{Z}'$ *such that* $f(x+iy)\to U$ *weakly in* $\mathcal{Z}'$ *as* $y\to\bar{0}$, $y\in C$; *and there exists a unique element* $V\in\mathcal{D}'$ *having support in* $S_A=\{t\in\mathbf{R}^n: u_C(t)\leq A\}$ *such that* $U=\mathcal{F}[V]$ *in* $\mathcal{Z}'$.

COROLLARY 4.1. *Let* $f(z)\in H(0;C)$ *where* C *is an open connected cone. There exist unique elements* $U\in\mathcal{Z}'$ *and* $V\in\mathcal{D}'$ *such that* $f(x+iy)\to U$ *weakly in* $\mathcal{Z}'$ *as* $y\to\bar{0}$, $y\in C$; $\operatorname{supp}(V)\subseteq C^*=S_0$; $U=\mathcal{F}[V]$ *in* $\mathcal{Z}'$; *and* $f(z)=\langle V,\exp(2\pi i\langle z,t\rangle)\rangle$, $z\in T^C$.

Proof. All results follow from Theorem 4.3 except for the representation of $f(z)$. In the proof of Theorem 4.3 for $A=0$, V is constructed to be the distributional derivative of a continuous function $g(t)$ on $\mathbf{R}^n$ which has support in C^* and which satisfies (4.1);

and $\operatorname{supp}(V)=\operatorname{supp}(g)\subseteq C^*$ as distributions since C^* is a regular set [30, pp. 98–99]. By Theorem 4.1, $(\exp(-2\pi\langle y,t\rangle)g(t))\in L^1\cap L^2$, $y\in C$. For $V=D^\gamma(g(t))$, γ being some n-tuple of nonnegative integers, the construction of the proof of Theorem 4.3 yields (see the similar step [6, (6), p. 846])

$$f(z)=z^\gamma\mathcal{F}\left[e^{-2\pi\langle y,t\rangle}g(t);x\right],\qquad z=x+iy\in T^C,$$

and this Fourier transform can be interpreted in both the L^1 and L^2 sense now. But by Theorem 4.2, $\langle V,\exp(2\pi i\langle z,t\rangle)\rangle\in H(0;C)$ for our present $V=D^\gamma(g(t))$. Thus the computation [3, (7.14), p. 1056], which is valid under the properties of $g(t)$ here for $z\in T^C$, together with the above representation of $f(z)$, $z\in T^C$, prove $f(z)=\langle V,\exp(2\pi i\langle z,t\rangle)\rangle$, $z\in T^C$, as desired. (The properties of $g(t)$ here are such that this proof is exactly the proof of [3, Thm. 8.2].) The proof is complete.

THEOREM 4.4. ([6, Thm. 2, p. 847]) *Let $f(z)\in H(A;C)$ where $A\geq 0$ and C is an open connected cone. Let $f(x+iy)\to U$ weakly in $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$, where U is unique. Then $U\in\mathcal{S}'$; there exists a unique element $V\in\mathcal{S}'$ such that* $\operatorname{supp}(V)\subseteq S_A=\{t\in\mathbf{R}^n: u_C(t)\leq A\}$ *and $U=\mathcal{F}[V]$ in $\mathcal{S}'$; and $f(z)=\langle V,\exp(2\pi i\langle z,t\rangle)\rangle$, $z\in T^C$.*

Special cases of results like Theorem 4.4 have appeared in the literature of quantum field theory; see [34, p. 61].

In the next two theorems C will denote an open convex cone such that [10, property (C), p. 395] is satisfied by each compact subcone C' of C. Further, we assume that the n-tuple $\alpha=(\alpha_1,\cdots,\alpha_n)$ is such that $\alpha_j\geq 1, j=1,\cdots,n$.

THEOREM 4.5. ([10, Thm. 1, p. 402]) *Let $f(z)\in H(A;C)$, $A\geq 0$. There exists a unique element $V\in\mathcal{S}'_\alpha$ with* $\operatorname{supp}(V)\subseteq S_A=\{t\in\mathbf{R}^n: u_C(t)\leq A\}$ *such that* $(\exp(-2\pi\langle y,t\rangle)V)\in\mathcal{S}'_\alpha$ *for all $y\in C$ and $f(x+iy)\to\mathcal{F}[V]\in(\mathcal{S}^\alpha)'$ in the weak topology of $(\mathcal{S}^\alpha)'$ as $y\to\bar{0}$, $y\in C$.*

THEOREM 4.6. ([10, Thm. 2, pp. 405–406]) *Let $f(z)\in H(0;C)$. There exists a unique element $V\in\mathcal{S}'_\alpha$ with* $\operatorname{supp}(V)\subseteq C^*=S_0$ *such that* $(\exp(-2\pi\langle y,t\rangle)V)\in\mathcal{S}'_\alpha$ *for all $y\in C$;*

$$f(z)=\langle V,e^{2\pi i\langle z,t\rangle}\rangle,\qquad z\in T^C; \tag{4.3}$$

$$f(z)=\mathcal{F}\left[e^{-2\pi\langle y,t\rangle}V\right],\qquad z=x+iy\in T^C, \tag{4.4}$$

where (4.4) holds as an equality in $(\mathcal{S}^\alpha)'$;

$$\{f(x+iy): y\in C, |y|\leq M\} \text{ is a strongly bounded set in } (\mathcal{S}^\alpha)' \text{ where } M \text{ is an arbitrary but fixed positive real number;} \tag{4.5}$$

$$f(x+iy)\to\mathcal{F}[V]\in(\mathcal{S}^\alpha)' \text{ in the weak topology of } (\mathcal{S}^\alpha)' \quad \text{as } y\to\bar{0},\quad y\in C. \tag{4.6}$$

As we noted earlier we have similarly generalized the boundary value results of [7, §§4 and 5] to the $H(A;C)$ functions in [3, §8]. Using analysis founded upon our analysis of [10], Pathak [26] has obtained results like those of [10] for the generalized function spaces W'_M and $(W^\Omega)'$ and for the class of holomorphic functions $H(A;C)$. (The space U_C^A of [26, p. 236] is identical to $H(A;C)$ for open connected (and hence convex) cones C and $A\geq 0$. Further, [26, Thm. 2, (ii) and (iii), p. 238] and [26, Thm. 3, (i) and (ii), p. 240] hold for $z\in T^C$.)

In addition to the growth (4.2) on holomorphic functions in tubes T^C we shall also be concerned in this paper with another growth which we introduce now. Let C be an open connected cone. For each compact subcone C' of C let $f(z)$ satisfy

$$|f(x+iy)|\leq M(C')(1+|z|)^N|y|^{-k},\qquad z=x+iy\in T^{C'}, \tag{4.7}$$

where $M(C')$ is a constant which depends on C' and where $N \geq 0$ and $k \geq 0$ are real numbers which are independent of C' and depend only on C and of course on the function. The growth (4.7) is of interest to us because any function $f(z)$ which is holomorphic in T^C and which satisfies (4.7) has a distributional boundary value in the weak and strong topologies of $\mathcal{S}'$ [36, p. 235]. We note that results of the type [36, p. 235] were first proved by Tillmann [35] for generalized half planes. Meise [24], [25] has extended the results of Tillmann to vector valued tempered distributions, and in [12] we have extended the result [36, p. 235] for functions holomorphic in tubes to the vector valued case. Note that any function $f(z)$ which satisfies (4.7) also satisfies (4.2) with $A=0$. Even with $A=0$ (4.2) is a more general growth than (4.7) because there are holomorphic functions which satisfy (4.2) but do not obtain $\mathcal{S}'$ boundary values and hence do not satisfy (4.7). We shall say more about this at the beginning of §6 below.

We conclude this section with a technical calculation which we need later. Let C be an open connected cone which is contained in or is any of the 2^n quadrants C_μ in $\mathbf{R}^n$. Put $F(z)=f(z)/X(z)$, $z \in T^C$, where $f(z)$ satisfies (4.7) and $X(z)$ is defined in (2.2) with the N being the N of (4.7). Let C' be any compact subcone of C and let $\delta>0$. If $y \in C' \subset C \subseteq C_\mu$ and $|y|>\delta$ then for each $j=1,\cdots,n$ there exists $\delta_j>0$ such that $((-1)^{\mu_j} y_j) \geq \delta_j > 0$. Thus from (2.5) and (4.7) we have for any compact subcone C' of C and any $\delta>0$ that

$$
\begin{aligned}
|F(z)| = \left|\frac{f(z)}{X(z)}\right| &\leq M'(C',\delta)\frac{(1+|z|)^N}{\Pi_{j=1}^n (1+|z_j|)^{N+n+2}} \\
&\leq M'(C',\delta)(1+|z|)^{-n-2}
\end{aligned} \tag{4.8}
$$

for all $z=x+iy \in T(C';\delta)=\mathbf{R}^n+i(C' \setminus (C' \cap N(\bar{0},\delta)))$ where $M'(C',\delta)$ is a constant which depends on C' and on δ and which is given by

$$
M'(C',\delta)=\frac{M(C')\delta^{-k}}{\Pi_{j=1}^n \left(K_j(\delta_j)\right)^{N+n+2}};
$$

here $M(C')$ is the constant from (4.7) and the $K_j(\delta_j)$, $j=1,\cdots,n$, are the constants from (2.5).

5. Holomorphic functions which have distributional boundary values and which are H^p functions. In this section and the next we obtain results like [28, Thm. 2] for functions holomorphic in tubes T^C in $\mathbf{C}^n$ and for various distribution spaces. We give necessary and sufficient conditions for functions which are holomorphic in tubes and which are known to have distributional boundary values to be H^p functions. The more interesting direction of our results in which the holomorphicity and growth are assumed and the H^p property obtained is proved first for tubes T^C where C is a cone that is contained in or is any quadrant C_μ in $\mathbf{R}^n$, and then we use this setting to prove the results for C being a general cone.

Because of the growths (3.20) and (3.17), the $\mathcal{S}'$ boundary value results of Vladimirov [36, p. 235], Tillmann [35], and others, and the importance of the tempered distributions $\mathcal{S}'$ in applications, the growth (4.7) is a natural growth to consider in obtaining results of the type which we desire. We consider the growth (4.7) in this section. In §6 we indicate results of the type which we obtain in this section but for the growth (4.2) with $A=0$, which is the growth that extends [28, (3.2)] to arbitrary tubes T^C in $\mathbf{C}^n$.

Let C be any open convex cone in $\mathbf{R}^n$ such that $\bar{C}$ does not contain any entire straight line throughout this section. Let $f(z)$ be holomorphic in $T^C=\mathbf{R}^n+iC$ and satisfy (4.7). By [36, p. 235] there is a unique $U\in\mathcal{S}'$ such that $f(x+iy)\to U$ in the weak topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$; and as noted before this convergence also holds in the strong topology of $\mathcal{S}'$ by [14, Cor. 8.4.9, p. 510] and the fact that $\mathcal{S}$ is a Montel space. We emphasize that this unique strong $\mathcal{S}'$ boundary value U of $f(x+iy)$ is obtained independently of the sequence $y\to\bar{0}$, $y\in C$, [36, p. 235], that is independently of the sequence $y\to\bar{0}$, $y\in C'\subset C$, for every compact subcone $C'\subset C$. We now begin our proof that if $U=h(x)\in L^p$, $1\le p\le\infty$, then $f(z)\in H^p(T^C)$.

THEOREM 5.1. *Let C be an open convex cone that is contained in or is any of the 2^n quadrants C_μ in $\mathbf{R}^n$. Let $f(z)$ be holomorphic in T^C and satisfy* (4.7). *Let the unique strong $\mathcal{S}'$ boundary value of $f(z)$, which exists, be $h(x)\in L^p$, $1\le p\le\infty$. Then $f(z)\in H^p(T^C)$, $1\le p\le\infty$, and*

$$f(z)=\int_{\mathbf{R}^n}h(t)Q(z;t)\,dt,\qquad z\in T^C. \tag{5.1}$$

Proof. Put

$$g_\varepsilon(z)=\frac{f(z)}{X_\varepsilon(z)},\qquad z\in T^C,\quad \varepsilon>0, \tag{5.2}$$

where

$$X_\varepsilon(z)=\prod_{j=1}^{n}\left(1-i\varepsilon(-1)^{\mu_j}z_j\right)^{N+n+2},\qquad \varepsilon>0, \tag{5.3}$$

with $\mu=(\mu_1,\cdots,\mu_n)$ being the n-tuple whose entries are 0 or 1 that defines C_μ, N being the constant of (4.7), and n being the dimension. By the proof of (2.4), $|1/X_\varepsilon(z)|\le 1$, $z\in T^{C_\mu}$, $\varepsilon>0$. Because of this, $g_\varepsilon(z)$ satisfies (4.7) since $f(z)$ does, and $g_\varepsilon(z)$ is holomorphic in T^C. By the argument in the paragraph preceding this theorem, for each $\varepsilon>0$ there is a unique $U_\varepsilon\in\mathcal{S}'$ such that

$$g_\varepsilon(x+iy)\to U_\varepsilon\quad\text{as } y\to\bar{0},\quad y\in C, \tag{5.4}$$

in the strong topology of $\mathcal{S}'$, and we emphasize again that this unique U_ε is independent of the sequence $y\to\bar{0}$, $y\in C$. For the present we let $\varepsilon>0$ be arbitrary but fixed. Let C' be an arbitrary compact subcone of $C\subseteq C_\mu$ and let $\delta>0$ be arbitrary. From the discussion of the last paragraph of §4 and the analysis of (4.8) we obtain the existence of a constant $M(C',\delta,\varepsilon)$ depending on C', δ, and ε such that

$$|g_\varepsilon(z)|\le M(C',\delta,\varepsilon)(1+|z|)^{-n-2},\qquad z\in T(C';\delta)=\mathbf{R}^n+i\big(C'\setminus(C'\cap N(\bar{0},\delta))\big). \tag{5.5}$$

Now put

$$G_\varepsilon(t)=\int_{\mathbf{R}^n}g_\varepsilon(x+iy)\exp(-2\pi i\langle x+iy,t\rangle)\,dx,\qquad y\in C,\quad t\in\mathbf{R}^n. \tag{5.6}$$

For any $y\in C$ there is a compact subcone C' of C and a $\delta>0$ such that $y\in(C'\setminus(C'\cap N(\bar{0},\delta)))$; thus (5.5) shows that $G_\varepsilon(t)$ is well defined as a function of $t\in\mathbf{R}^n$ for each $y\in C$. In fact $G_\varepsilon(t)$ is a continuous function of $t\in\mathbf{R}^n$ for $y\in C$; and because of the growth (5.5), the same type of analysis used in [6, pp. 846–847] to show the independence of $y\in C$ of the function in [6, (5), p. 846] and the support of this function to be in $S_A=\{t:u_C(t)\le A\}$ there can be used here to show that $G_\varepsilon(t)$ in (5.6) is independent of $y\in C$ and $\operatorname{supp}(G_\varepsilon)\subseteq C^*$ for each $\varepsilon>0$. (This same type of argument has also been used

in the proof of [3, Thm. 7.1, pp. 1055–1056].) For any compact subcone C' of C and any $\delta>0$, (5.5) yields

$$|G_\varepsilon(t)|\le M'(C',\delta,\varepsilon)e^{2\pi\langle y,t\rangle}, \qquad t\in\mathbf{R}^n, \quad y\in\big(C'\setminus\big(C'\cap N(\bar{0},\delta)\big)\big), \tag{5.7}$$

and (5.7) holds independently of $y\in(C'\setminus(C'\cap N(\bar{0},\delta)))$ since $G_\varepsilon(t)$ is independent of $y\in C$. Also from (5.5), $g_\varepsilon(x+iy)\in L^1\cap L^2$ as a function of $x\in\mathbf{R}^n$ for each $y\in C$. Thus (5.6) can be rewritten as

$$e^{-2\pi\langle y,t\rangle}G_\varepsilon(t)=\mathcal{F}^{-1}\big[g_\varepsilon(x+iy);t\big], \qquad y\in C, \tag{5.8}$$

where $\mathcal{F}^{-1}$ can be interpreted as either the L^1 or L^2 inverse Fourier transform. By the Plancherel theory we have $(\exp(-2\pi\langle y,t\rangle)G_\varepsilon(t))\in L^2$, $y\in C$, and

$$g_\varepsilon(x+iy)=\mathcal{F}\big[e^{-2\pi\langle y,t\rangle}G_\varepsilon(t);x\big], \qquad z=x+iy\in T^C, \tag{5.9}$$

with this Fourier transform being in the L^2 sense.

Since $G_\varepsilon(t)$ is continuous, $\operatorname{supp}(G_\varepsilon)\subseteq C^*$, and $G_\varepsilon(t)$ satisfies (5.7) independently of $y\in(C'\setminus(C'\cap N(\bar{0},\delta)))$ for any compact subcone C' of C and any $\delta>0$, then by Theorem 4.1 we have $(\exp(-2\pi\langle y,t\rangle)G_\varepsilon(t))\in L^p$ for all p, $1\le p<\infty$, as a function of $t\in\mathbf{R}^n$ for each fixed $y\in C$. Thus the Fourier transform in (5.9) can also be interpreted in the L^1 sense, and (5.9) becomes

$$g_\varepsilon(x+iy)=\int_{\mathbf{R}^n}G_\varepsilon(t)e^{2\pi i\langle z,t\rangle}\,dt, \qquad z\in T^C, \tag{5.10}$$

with the right side being a holomorphic function of $z\in T^C$ since the left side is. (The right side of (5.10) is also holomorphic in T^C because of Theorem 4.2.)

Both $G_\varepsilon(t)$ and $(\exp(-2\pi\langle y,t\rangle)G_\varepsilon(t))$, $y\in C$, are elements of $\mathcal{D}'$. Also $g_\varepsilon(x+iy)\in\mathcal{Z}'$ as a function of $x\in\mathbf{R}^n$ for each $y\in C$. Thus (5.9) holds as an equality in $\mathcal{Z}'$. Let $\phi\in\mathcal{D}$ and $\psi(x)=\mathcal{F}[\phi(t);x]\in\mathcal{Z}$. Arguing exactly as in [36, p. 237, ll. 10–14] or [6, (11), p. 847] and using the Fourier transform from $\mathcal{D}'$ to $\mathcal{Z}'$ we have

$$\langle g_\varepsilon(x+iy),\psi(x)\rangle=\langle e^{-2\pi\langle y,t\rangle}G_\varepsilon(t),\phi(t)\rangle\to\langle G_\varepsilon(t),\phi(t)\rangle=\langle\mathcal{F}\big[G_\varepsilon(t)\big],\psi\rangle \tag{5.11}$$

as $y\to\bar{0}$, $y\in C$, which proves that

$$g_\varepsilon(x+iy)\to\mathcal{F}\big[G_\varepsilon(t)\big] \quad \text{as } y\to\bar{0}, \quad y\in C, \tag{5.12}$$

in the weak topology of $\mathcal{Z}'$ with $\mathcal{F}[G_\varepsilon(t)]$ being the Fourier transform of $G_\varepsilon(t)$ from $\mathcal{D}'$ to $\mathcal{Z}'$. But (5.4) holds in the strong and weak topology of $\mathcal{S}'$; and since $U_\varepsilon\in\mathcal{S}'\subset\mathcal{Z}'$ and the $\mathcal{S}'$ topology is stronger than that of $\mathcal{Z}'$, we have that (5.4) also holds in the weak topology of $\mathcal{Z}'$ and $U_\varepsilon=\mathcal{F}[G_\varepsilon(t)]$ in $\mathcal{Z}'$. Thus $G_\varepsilon(t)=\mathcal{F}^{-1}[U_\varepsilon]$ in $\mathcal{D}'$. But $U_\varepsilon\in\mathcal{S}'\subset\mathcal{Z}'$; so that $\mathcal{F}^{-1}[U_\varepsilon]\in\mathcal{S}'\subset\mathcal{D}'$. Since $G_\varepsilon(t)=\mathcal{F}^{-1}[U_\varepsilon]$ in $\mathcal{D}'$ and $\mathcal{F}^{-1}[U_\varepsilon]\in\mathcal{S}'\subset\mathcal{D}'$ we thus have that $G_\varepsilon(t)$ can be extended to be an element of $\mathcal{S}'$ with $G_\varepsilon(t)=\mathcal{F}^{-1}[U_\varepsilon]\in\mathcal{S}'$ in $\mathcal{S}'$ ([36, pp. 237–238, especially p. 238, lines 1–6], [6, proof of Thm. 2, p. 847], or [12, p. 330, lines 10–20]); and hence

$$U_\varepsilon=\mathcal{F}\big[G_\varepsilon(t)\big]\in\mathcal{S}' \quad \text{in } \mathcal{S}'. \tag{5.13}$$

(Once we have that $G_\varepsilon(t)\in\mathcal{S}'$ above, (5.13) also follows in another way: $(\exp(-2\pi\langle y,t\rangle)G_\varepsilon(t))\in L^2\subset\mathcal{S}'$, $y\in C$, and $g_\varepsilon(x+iy)\in\mathcal{S}'$ as a function of $x\in\mathbf{R}^n$ for $y\in C$; thus (5.9) holds as an equality in $\mathcal{S}'$. We can now prove directly using analysis as in (5.11) that $g_\varepsilon(x+iy)\to\mathcal{F}[G_\varepsilon(t)]\in\mathcal{S}'$ in the weak, and hence strong, topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$. But from (5.4), $U_\varepsilon\in\mathcal{S}'$ is the unique $\mathcal{S}'$ boundary value of $g_\varepsilon(x+iy)$; hence we must have (5.13).)

Now $X_\varepsilon(x+iy)$ is a multiplier in $\mathcal{S}'$ as a function of $x\in\mathbf{R}^n$ for each fixed $y\in C$. Also

$$X_\varepsilon(x+iy)\to X_\varepsilon(x) \quad \text{as } y\to\bar{0},\quad y\in C, \tag{5.14}$$

in the strong topology of $\mathcal{S}'$ by the proof of Lemma 2.2, and $X_\varepsilon(x)$ is a multiplier in $\mathcal{S}'$. From (5.2), (5.4), and (5.14) we obtain

$$\lim_{\substack{y\to\bar{0}\\ y\in C}} f(x+iy)=\lim_{\substack{y\to\bar{0}\\ y\in C}} X_\varepsilon(x+iy)g_\varepsilon(x+iy)=X_\varepsilon(x)U_\varepsilon \tag{5.15}$$

in the strong topology of $\mathcal{S}'$. But by hypothesis, the unique strong $\mathcal{S}'$ boundary value of $f(z)$, which exists, is $h(x)\in L^p$, $1\le p\le\infty$. We thus have

$$X_\varepsilon(x)U_\varepsilon=h(x) \quad \text{in } \mathcal{S}'. \tag{5.16}$$

The proof of Lemma 2.2 and (5.16) now yield

$$U_\varepsilon=\frac{h(x)}{X_\varepsilon(x)} \quad \text{in } \mathcal{S}'. \tag{5.17}$$

The proof will now proceed by considering different values for p. First assume $1\le p\le 2$. Our goal is to construct a function which we know is in $H^p(T^C)$, $1\le p\le 2$, and then to prove that it equals $f(z)$. Since $h(x)\in L^p$, $1\le p\le 2$, and $|1/X_\varepsilon(x)|\le 1$, $\varepsilon>0$, then $(h(x)/X_\varepsilon(x))\in L^p$, $1\le p\le 2$. Recall that $G_\varepsilon(t)$ is a continuous function of $t\in\mathbf{R}^n$ with $\operatorname{supp}(G_\varepsilon)\subseteq C^*$; thus because of (5.13) and (5.17), we have exactly the situation described in the last paragraph of §2. By the discussion of that paragraph and by the calculations (2.9) and (2.10), there exists a function $H_\varepsilon(t)=\mathcal{F}^{-1}[h(x)/X_\varepsilon(x);t]$ which is in L^∞ if $p=1$ and in L^q, $(1/p)+(1/q)=1$, if $1<p\le 2$, such that $\operatorname{supp}(H_\varepsilon)\subseteq C^*$ almost everywhere and

$$\begin{aligned}\mathcal{F}[G_\varepsilon(t)]=U_\varepsilon=\frac{h(x)}{X_\varepsilon(x)}=\mathcal{F}[H_\varepsilon(t)],\\ G_\varepsilon(t)=\mathcal{F}^{-1}[U_\varepsilon]=\mathcal{F}^{-1}[h(x)/X_\varepsilon(x)]=H_\varepsilon(t)\end{aligned} \tag{5.18}$$

as equalities in $\mathcal{S}'$ and

$$\int_{C^*} G_\varepsilon(t)e^{2\pi i\langle z,t\rangle}\,dt=\int_{C^*} H_\varepsilon(t)e^{2\pi i\langle z,t\rangle}\,dt, \qquad z\in T^C. \tag{5.19}$$

By (5.10), the facts that $\operatorname{supp}(G_\varepsilon)\subseteq C^*$ and $\operatorname{supp}(H_\varepsilon)\subseteq C^*$ almost everywhere, (5.19), the fact that $H_\varepsilon(t)=\mathcal{F}^{-1}[h(x)/X_\varepsilon(x);t]$ in the function sense where $(h(x)/X_\varepsilon(x))\in L^p$, $1\le p\le 2$, and the proof of Lemma 3.3 we have

$$g_\varepsilon(z)=\int_{C^*} G_\varepsilon(t)e^{2\pi i\langle z,t\rangle}\,dt=\int_{\mathbf{R}^n}\frac{h(t)}{X_\varepsilon(t)}K(z-t)\,dt, \qquad z\in T^C, \tag{5.20}$$

where $K(z-t)$ is the Cauchy kernel.

Now let w be an arbitrary but fixed point of T^C and consider the function $(K(z+w)g_\varepsilon(z))$, $z\in T^C$. This product is holomorphic in $z\in T^C$ by Lemma 3.2 and the fact that $g_\varepsilon(z)$ is. Recall that $g_\varepsilon(z)$ defined in (5.2) satisfies exactly the growth (4.7) of $f(z)$ since $|1/X_\varepsilon(z)|\le 1$, $\varepsilon>0$, $z\in T^C$. Thus by (3.6), $(K(z+w)g_\varepsilon(z))$ satisfies the growth obtained by multiplying the growth of $g_\varepsilon(z)$ by the constant M_v, $v=\operatorname{Im}(w)$, of (3.6), which is independent of $z\in T^C$. By Lemma 3.2, $K(x+iy+w)\to K(x+w)$ in the strong topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$. Thus by this convergence, the existence of U_ε in

(5.4), and (5.17) we have

$$\lim_{\substack{y\to 0\\ y\in C}} K(x+iy+w)g_\varepsilon(x+iy)=K(x+w)U_\varepsilon=K(x+w)\frac{h(x)}{X_\varepsilon(x)} \tag{5.21}$$

in the strong topology of $\mathcal{S}'$; and $(K(x+w)(h(x)/X_\varepsilon(x)))\in L^p$, $1\le p\le 2$, since $h(x)\in L^p$, $1\le p\le 2$, here and both $K(x+w)$ and $1/X_\varepsilon(x)$ are bounded for $x\in\mathbf{R}^n$. We thus have exactly the same situation and type of hypothesis with respect to $(K(z+w)g_\varepsilon(z))$, $z\in T^C$, as we did with $g_\varepsilon(z)$ in obtaining (5.20) from the beginning of the proof of this theorem. Thus by (5.21) and the holomorphicity and growth of $(K(z+w)g_\varepsilon(z))$ for $z\in T^C$ and by exactly the same proof as in obtaining (5.20) we have

$$K(z+w)g_\varepsilon(z)=\int_{\mathbf{R}^n}\frac{h(t)}{X_\varepsilon(t)}K(t+w)K(z-t)\,dt,\qquad z\in T^C. \tag{5.22}$$

This equality holds for w being arbitrary but fixed in T^C. For $z=x+iy\in T^C$ in (5.22) we now choose $w=-x+iy\in T^C$. With this choice of w, $(K(t+w)K(z-t))=|K(z-t)|^2$ and $K(z+w)=K(2iy)$; and (5.22) yields

$$g_\varepsilon(z)=\int_{\mathbf{R}^n}\frac{h(t)}{X_\varepsilon(t)}Q(z;t)\,dt,\qquad z\in T^C, \tag{5.23}$$

where $Q(z;t)$ is the Poisson kernel corresponding to T^C. (Our argument in this paragraph leading to (5.23) is an adaptation of the argument in the proof of [33, Thm. 3.9, p. 106].) The Poisson integral on the right of (5.23) is holomorphic in $z\in T^C$ since $g_\varepsilon(z)$ is. Combining (5.20) and (5.23) we thus have

$$\begin{aligned} g_\varepsilon(z)=\int_{C^*}G_\varepsilon(t)e^{2\pi i\langle z,t\rangle}\,dt&=\int_{\mathbf{R}^n}\frac{h(t)}{X_\varepsilon(t)}K(z-t)\,dt\\ &=\int_{\mathbf{R}^n}\frac{h(t)}{X_\varepsilon(t)}Q(z;t)\,dt,\qquad z\in T^C, \end{aligned} \tag{5.24}$$

and all four functions of $z\in T^C$ in (5.24) are holomorphic in T^C.

Since $|1/X_\varepsilon(x)|\le 1$ for all $x\in\mathbf{R}^n$ and all $\varepsilon>0$, then

$$\left|\frac{h(x)}{X_\varepsilon(x)}-h(x)\right|^p\le 2^p\left(\left|\frac{h(x)}{X_\varepsilon(x)}\right|^p+|h(x)|^p\right)\le 2^{p+1}|h(x)|^p$$

and the right side is in L^1 since $h\in L^p$, $1\le p\le 2$, here and is independent of $\varepsilon>0$. By the Lebesgue dominated convergence theorem

$$\lim_{\varepsilon\to 0+}\int_{\mathbf{R}^n}\left|\frac{h(x)}{X_\varepsilon(x)}-h(x)\right|^p dx=0,\qquad 1\le p\le 2. \tag{5.25}$$

Now put

$$G(z)=\int_{\mathbf{R}^n}h(t)Q(z;t)\,dt,\qquad z\in T^C. \tag{5.26}$$

(We shall show that $G(z)\in H^p(T^C)$, $1\le p\le 2$, and then prove $G(z)=f(z)$, $z\in T^C$.) Let z_0 be an arbitrary but fixed point in T^C. Since C is open we can choose a closed neighborhood $N(z_0,\delta)=\{z:|z-z_0|\le\delta,\ \delta>0\}$ of z_0 contained in T^C. If $1<p\le 2$ we use

(5.23), (5.26), and Hölder's inequality to obtain

$$|g_\varepsilon(z)-G(z)|\leq\left\|\frac{h(t)}{X_\varepsilon(t)}-h(t)\right\|_{L^p}\|Q(z;t)\|_{L^q},\qquad z\in T^C, \tag{5.27}$$

$(1/p)+(1/q)=1$; and if $p=1$ we use (5.23), (5.26), and (3.12) to yield

$$|g_\varepsilon(z)-G(z)|\leq\frac{(K(iy))^2}{K(2iy)}\int_{\mathbf{R}^n}\left|\frac{h(t)}{X_\varepsilon(t)}-h(t)\right|dt,\qquad z\in T^C. \tag{5.28}$$

Thus (5.25), (5.27), and (5.28) together with Lemma 3.4 and its proof yield that $g_\varepsilon(z)\to G(z)$ uniformly in z for $z\in\{z:|z-z_0|<\delta\}$ as $\varepsilon\to 0+$. From this fact and the fact that $g_\varepsilon(z)$ is holomorphic in T^C for each $\varepsilon>0$, we conclude that $G(z)$ is holomorphic at $z_0\in T^C$ and hence in the whole of T^C since z_0 is an arbitrary point of T^C. We now apply Lemma 3.5 to $G(z)$, which is defined by (5.26), and conclude that $G(z)\in H^p(T^C)$, $1\leq p\leq 2$, since $h(t)\in L^p$, $1\leq p\leq 2$, here.

For $1\leq p\leq 2$ and $\phi\in\mathcal{S}$ we have by Hölder's inequality that

$$|\langle G(x+iy),\phi(x)\rangle-\langle h(x),\phi(x)\rangle|\leq\|G(x+iy)-h(x)\|_{L^p}\|\phi\|_{L^q}. \tag{5.29}$$

Inequality (5.29) together with the definition (5.26) of $G(z)$ and [22, Prop. 3(c)], the fact that $G(x+iy)\to h(x)$ in L^p as $y\to\bar{0}$, $y\in C$, prove that $G(x+iy)\to h(x)$ in the weak topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$; hence this convergence is in the strong topology of $\mathcal{S}'$ also as we have argued before.

We now have that $(f(z)-G(z))$ is a holomorphic function of $z\in T^C$. By hypothesis, $f(z)$ satisfies (4.7) and hence satisfies (4.2) with $A=0$. Since $G(z)\in H^p(T^C)$, $1\leq p\leq 2$, then $G(z)$ satisfies (3.17) by Theorem 3.1; hence $G(z)$ satisfies (4.2) with $A=0$. Thus $(f(z)-G(z))\in H(0;C)$ as defined in §4. Further, both $f(z)$ and $G(z)$ converge to $h(x)\in L^p\subset\mathcal{S}'$ in the strong topology of $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$; so that

$$(f(x+iy)-G(x+iy))\to 0\quad\text{as } y\to\bar{0},\quad y\in C, \tag{5.30}$$

in the strong (and weak) topology of $\mathcal{S}'$ with the boundary value $U=0\in\mathcal{S}'$ being unique in (5.30). By Theorem 4.4 there exists a unique element $V\in\mathcal{S}'$ with $\operatorname{supp}(V)\subseteq C^*$ such that $0=\mathcal{F}[V]$ in $\mathcal{S}'$ and

$$f(z)-G(z)=\langle V,e^{2\pi i\langle z,t\rangle}\rangle=\langle V,\lambda(t)e^{2\pi i\langle z,t\rangle}\rangle,\qquad z\in T^C, \tag{5.31}$$

where $\lambda(t)$ is the function of (2.8) with $b=0$ there and $(\lambda(t)\exp(2\pi i\langle z,t\rangle))\in\mathcal{S}$ for $z\in T^C$. But the Fourier transform on $\mathcal{S}'$ maps $\mathcal{S}'$ one to one and onto $\mathcal{S}'$ as does the inverse Fourier transform on $\mathcal{S}'$. Thus $0=\mathcal{F}[V]$ in $\mathcal{S}'$ implies $V=\mathcal{F}^{-1}[0]=0$ in $\mathcal{S}'$. This together with (5.31) proves that $f(z)=G(z)$, $z\in T^C$. The proof of our result for the cases $1\leq p\leq 2$ is thus complete since we already know that $G(z)\in H^p(T^C)$, $1\leq p\leq 2$, and (5.1) holds by the definition (5.26) of $G(z)$.

We now complete the proof of Theorem 5.1 by proving the result for the remaining cases that $2<p\leq\infty$. Recall that our analysis from the beginning of the proof through (5.17) holds for $1\leq p\leq\infty$. For any $\varepsilon>0$

$$\left|\frac{1}{X_\varepsilon(x)}\right|=\varepsilon^{-n(N+n+2)}\prod_{j=1}^{n}\left(\varepsilon^{-2}+x_j^2\right)^{-1-N/2-n/2},$$

which shows that for each $\varepsilon>0$, $(1/X_\varepsilon(x))\in L^q$ for all q, $1\leq q\leq\infty$. By hypothesis in the present case, $h(x)\in L^p$, $2<p\leq\infty$. Thus $(h(x)/X_\varepsilon(x))\in L^1\cap L^p$, $2<p\leq\infty$. Now if $p=\infty$ then $(h(x)/X_\varepsilon(x))\in L^2$ since $h\in L^\infty$ and $(1/X_\varepsilon(x))\in L^2$. If $2<p<\infty$, analysis

as in (3.25) proves $(h(x)/X_\varepsilon(x))\in L^2$ here. We thus have $(h(x)/X_\varepsilon(x))\in L^1\cap L^2\cap L^p$, $2<p\leq\infty$, and (5.17) holds. Since $(h(x)/X_\varepsilon(x))\in L^2$ we thus may use exactly the proof given above for the case $p=2$ to conclude that (5.18), (5.19), (5.20), (5.23), and (5.24) hold such that all four functions of $z\in T^C$ in (5.24) are holomorphic in T^C for each $\varepsilon>0$. If $2<p<\infty$, (5.25) holds here by exactly the same analysis; hence for the Poisson integral representation of $g_\varepsilon(z)$ in (5.23) and (5.24) and the definition of $G(z)$ in (5.26), the exact same analysis as in the case $1<p\leq 2$ now proves for our present case of $2<p<\infty$ that $g_\varepsilon(z)\to G(z)$ uniformly in z for z in an open neighborhood of each fixed point $z_0\in T^C$ as $\varepsilon\to 0+$. For the case $p=\infty$, we choose a closed neighborhood contained in T^C about each fixed point $z_0\in T^C$, which can be done since C is open, and apply the proof of Lemma 3.6 to yield that $g_\varepsilon(z)\to G(z)$ uniformly on the closed neighborhood of $z_0\in T^C$ and hence on the corresponding open neighborhood about $z_0\in T^C$ as $\varepsilon\to 0+$. Thus for any p, $2<p\leq\infty$, we conclude as before that $G(z)$ is holomorphic at each point $z_0\in T^C$ and hence in the whole of T^C since $g_\varepsilon(z)$ is holomorphic in T^C for each $\varepsilon>0$. By Lemma 3.5 we now have $G(z)\in H^p(T^C)$, $2<p\leq\infty$. Using [22, Prop. 3(c) and 3(d)], the inequality (5.29) if $2<p<\infty$, and the definition of the weak-star topology of L^∞ if $p=\infty$, we obtain that $G(x+iy)\to h(x)$ weakly and hence strongly in $\mathcal{S}'$ as $y\to\bar{0}$, $y\in C$. The proof for $2<p\leq\infty$ is completed by exactly the same argument on $(f(z)-G(z))$ as at the conclusion of the proof for $1\leq p\leq 2$. We have $(f(z)-G(z))\in H(0;C)$ and (5.30) and (5.31) hold with $V=0$ in (5.31). Thus $f(z)=G(z)\in H^p(T^C)$, $2<p\leq\infty$, and (5.1) holds because of the definition of $G(z)$ in (5.26). The proof of Theorem 5.1 is complete.

We note an error in the proof of [28, Thm. 2]. The statement "$g_\varepsilon(x)\to f(x)$ in $L^p(-\infty,\infty)$ as $\varepsilon\to 0$" for the case $p=\infty$ in [28, p. 517, line 20] is false. $f(x)$ there could be $f(x)=1$, $-\infty<x<\infty$, for example in the case $p=\infty$; then $g_\varepsilon(x)=(f(x)/\phi_\varepsilon(x))\to f(x)$ in $L^\infty(-\infty,\infty)$ as $\varepsilon\to 0+$ if and only if $(1/\phi_\varepsilon(x))\to 1$ uniformly on $-\infty<x<\infty$ as $\varepsilon\to 0+$ where $\phi_\varepsilon(x)$ is defined in [28, (3.4), p. 516]; and this latter convergence is false. For $p=\infty$ a correct approach is to use the proof of our Lemma 3.6 to achieve the desired result at this step in the proof of [28, Thm. 2].

We now extend Theorem 5.1 to arbitrary tubes of the type which we are considering in this paper.

THEOREM 5.2. *Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. Let $f(z)$ be holomorphic in T^C and satisfy* (4.7). *Let the unique strong $\mathcal{S}'$ boundary value of $f(z)$, which exists, be $h(x)\in L^p$, $1\leq p\leq\infty$. Then $f(z)\in H^p(T^C)$, $1\leq p\leq\infty$, and* (5.1) *holds.*

Proof. For each of the 2^n quadrants C_μ consider $C\cap C_\mu$. Let S_j, $j=1,\cdots,k$, be an enumeration of the intersections $C\cap C_\mu$ which are nonempty; then each S_j is an open convex cone which is contained in a quadrant in $\mathbf{R}^n$. Put

$$f_j(z)=f(z),\qquad z\in T^{S_j}=\mathbf{R}^n+iS_j,\quad j=1,\cdots,k. \tag{5.32}$$

Then each $f_j(z)$ satisfies the hypothesis of Theorem 5.1 for $z\in T^{S_j}$, $j=1,\cdots,k$, with respect to holomorphicity and growth. From the discussion in the paragraph immediately preceding Theorem 5.1, $f(z)$ obtains its strong $\mathcal{S}'$ boundary value independently of how $y\to\bar{0}$, $y\in C$, and this boundary value is $h(x)\in L^p$, $1\leq p\leq\infty$, here. Thus each $f_j(z)$ has $h(x)$ as its unique strong $\mathcal{S}'$ boundary value as $y\to\bar{0}$, $y\in S_j$, $j=1,\cdots,k$. By Theorem 5.1, we obtain $f_j(z)\in H^p(T^{S_j})$, $j=1,\cdots,k$, and

$$f_j(z)=\int_{\mathbf{R}^n}h(t)Q(z;t)\,dt,\qquad z\in T^{S_j},\quad j=1,\cdots,k. \tag{5.33}$$

For $1 \leq p < \infty$, there exist constants $A_j, j = 1, \cdots, k$, such that

$$\int_{\mathbf{R}^n} |f_j(x+iy)|^p \, dx \leq A_j^p, \qquad y \in S_j, \quad j = 1, \cdots, k, \tag{5.34}$$

and each A_j is independent of $y \in S_j$. Put

$$A = \max\{A_1, A_2, \cdots, A_k\}. \tag{5.35}$$

Now let $y \in C$ such that $y \notin S_j$, $j = 1, \cdots, k$. Then y is on the topological boundary of some S_j. Choose a sequence of points $\{y_{j,n}\} \subset S_j$ such that $y_{j,n} \to y$ as $n \to \infty$. By the fact that $f(z)$ is holomorphic in T^C, the definition (5.32) of $f_j(z)$, $j = 1, \cdots, k$, Fatou's lemma, (5.34), and (5.35) we have for $1 \leq p < \infty$ and $y \in C$ such that $y \notin S_j$, $j = 1, \cdots, k$, that

$$\int_{\mathbf{R}^n} |f(x+iy)|^p \, dx \leq \liminf_{n \to \infty} \int_{\mathbf{R}^n} |f(x+iy_{j,n})|^p \, dx \leq A_j^p \leq A^p. \tag{5.36}$$

Combining (5.32), (5.34), (5.35) and (5.36) we have for $1 \leq p < \infty$ that

$$\int_{\mathbf{R}^n} |f(x+iy)|^p \, dx \leq A^p, \qquad y \in C,$$

where A is independent of $y \in C$. Thus $f(z) \in H^p(T^C)$, $1 \leq p < \infty$, as desired.

If $p = \infty$ we apply Theorem 5.1 and obtain $f_j(z) \in H^\infty(T^{S_j})$, $j = 1, \cdots, k$; hence

$$|f_j(z)| \leq B_j, \qquad z \in T^{S_j}, \quad j = 1, \cdots, k, \tag{5.37}$$

for positive constants $B_j, j = 1, \cdots, k$, which are independent of $z \in T^{S_j}$. Put

$$B = \max\{B_1, B_2, \cdots, B_k\}. \tag{5.38}$$

Again let $y \in C$ such that $y \notin S_j$, $j = 1, \cdots, k$, and choose a sequence $\{y_{j,n}\}$ in an appropriate S_j which converges to y as $n \to \infty$. Since $f(z)$ is holomorphic and hence continuous at $z = x + iy$, a simple continuity argument together with (5.32), (5.37) and (5.38) proves that

$$|f(x+iy)| \leq 1 + B \tag{5.39}$$

for any $y \in C$ such that $y \notin S_j$, $j = 1, \cdots, k$. Then (5.32), (5.37), (5.38) and (5.39) prove that (5.39) holds for all $z \in T^C$ where B is independent of $z \in T^C$. Hence $f(z) \in H^\infty(T^C)$ in the case $p = \infty$.

We now have $f(z) \in H^p(T^C)$, $1 \leq p \leq \infty$. By [22, Prop. 4] there exists a function $H(t) \in L^p$ such that

$$f(z) = \int_{\mathbf{R}^n} H(t) Q(z;t) \, dt, \qquad z \in T^C. \tag{5.40}$$

But for $z \in T^{S_j}$, $j = 1, \cdots, k$, (5.33) holds. Thus by (5.32), (5.33), (5.40) and [22, Prop. 3(c) and 3(d)] we have that $h(t) = H(t)$ almost everywhere for $t \in \mathbf{R}^n$. Hence the conclusion (5.1) in this theorem follows from (5.40). The proof of Theorem 5.2 is complete.

As we noted at the beginning of this section we chose to consider the growth (4.7) here because it naturally extends the known growth of H^p functions in tubes as given in Theorem 3.1, because of the $\mathcal{S}'$ boundary value results of [36, p. 235] and [35] and the growths considered there, and because of the importance of $\mathcal{S}'$ in applications. We note for emphasis that Theorems 5.1 and 5.2 will hold for any growth which will satisfy the growth of [36, p. 235] and such that the technical construction in the proof of Theorem

5.1 will hold. In particular these theorems will hold for the functions [32, Thm. II.5, (II.11b), p. 56] and for those of Tillmann [35, (5), p. 110].

Two corollaries follow immediately from Theorem 5.2.

COROLLARY 5.1. *Let the cone C and the holomorphic function $f(z)$, $z\in T^C$, satisfy the hypotheses of Theorem* 5.2 *for* $1<p\leq 2$. *Then* (5.1) *holds and there exists a function* $g(t)\in L^q$, $(1/p)+(1/q)=1$, *with* $\operatorname{supp}(g)\subseteq C^*$ *almost everywhere such that*

$$f(z)=\int_{\mathbf{R}^n} g(t)e^{2\pi i\langle z,t\rangle}\,dt, \qquad z\in T^C. \tag{5.41}$$

If $p=2$ *we further have*

$$f(z)=\int_{\mathbf{R}^n} h(t)K(z-t)\,dt, \qquad z\in T^C, \tag{5.42}$$

where $h(x)\in L^2$ *is the unique strong* $\mathcal{S}'$ *boundary value of* $f(z)$ *in the hypothesis of Theorem* 5.2.

Proof. By Theorem 5.2, $f(z)\in H^p(T^C)$, $1<p\leq 2$, and (5.1) holds. The existence of $g(t)\in L^q$, $(1/p)+(1/q)=1$, with $\operatorname{supp}(g)\subseteq C^*$ almost everywhere such that (5.41) holds, now follows by Theorem 3.2, case III. If $p=2$, there exists a function $H(t)\in L^2$ such that

$$f(z)=\int_{\mathbf{R}^n} H(t)K(z-t)\,dt=\int_{\mathbf{R}^n} H(t)Q(z;t)\,dt, \qquad z\in T^C, \tag{5.43}$$

with $H(t)=\mathcal{F}[g(x);t]$ in L^2 [33, Thms. 3.6 and 3.9, pp. 103–106]. But (5.1) holds for the $h\in L^2$ from Theorem 5.2. By the same argument as at the end of the proof of Theorem 5.2, (5.1) and (5.43) imply that $h(t)=H(t)$ almost everywhere; hence (5.42) is proved by (5.43).

COROLLARY 5.2. *let the cone C and the holomorphic function $f(z)$, $z\in T^C$, satisfy the hypotheses of Theorem* 5.2 *with the unique strong* $\mathcal{S}'$ *boundary value of* $f(z)$, *which exists, being a constant* K. *Then* $f(z)=K$, $z\in T^C$.

Proof. By Theorem 5.2, (5.1) holds with $h(t)=K$. By (3.4) we then obtain $f(z)=K$, $z\in T^C$, as desired.

We now state a converse to Theorem 5.2.

THEOREM 5.3. *Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. Let* $f(z)\in H^p(T^C)$, $1\leq p\leq\infty$. *There exists* $h(x)\in L^p$ *such that*

$$f(x+iy)\to h(x) \quad \text{as } y\to\overline{0},\ y\in C, \tag{5.44}$$

in L^p *if* $1\leq p<\infty$ *and in the weak-star topology of* L^∞ *if* $p=\infty$; *and for all* p, $1\leq p\leq\infty$, (5.44) *holds in the strong topology of* $\mathcal{S}'$. *Further, if* $1\leq p<\infty$ *then* $f(z)$ *satisfies* (3.17) *and hence* (4.7); *if* $p=\infty$, $f(z)$ *is bounded on* T^C *and hence satisfies* (4.7); *and for all* p, $1\leq p\leq\infty$, (5.1) *holds.*

The proof of Theorem 5.3 is immediate using [22, Prop. 4 and Prop. 3(c) and 3(d)] and arguments that we have used previously in this paper. Details are left to the interested reader.

6. Results for other distribution spaces. In this section we note results like those of §5 for the distribution spaces which are associated with the space of functions $H(0;C)$ as defined in §4. The elements of $H(0;C)$ are defined by the growth (4.2) with $A=0$ there. As we previously noted, functions which satisfy (4.7), a growth which yields $\mathcal{S}'$ boundary values, also satisfy (4.2) with $A=0$. One naturally asks whether one actually obtains a different distributional boundary value theory by considering the more general growth (4.2). The answer is yes. There are holomorphic functions in tubes

which satisfy the growth (4.2), even for $A=0$, but which do not obtain $\mathcal{S}'$ distributional boundary values and hence can't satisfy (4.7); see [1, p. 306] and [34, (2–67), p. 54] for relevant examples. Thus obtaining distributional boundary value results for holomorphic functions in tubes which satisfy (4.2) with respect to the distribution spaces $\mathcal{Z}'$, K'_r, $r\geq 1$, $(\mathcal{S}^\alpha)'$, or $(W^\Omega)'$ yields information not obtainable by considering the growth (4.7) and can have importance in applications as inferred in [28, §1, last ¶ and Thm. 2] and [13].

Let C be an open convex cone. $f(z)\in H(0;C)$ obtains a unique weak distributional boundary value in $\mathcal{Z}'$ by Theorem 4.2 of this paper and in K'_r, $r\geq 1$, by [3, Thm. 8.1]; and in the case of K'_r the convergence of $f(x+iy)$ to a unique element of K'_r as $y\to 0$, $y\in C$, is in the strong topology of K'_r also [3, Thm. 8.1]. (The $H(A;C)$ functions defined in §4 are slightly more general than the $F_1(A;C)$ functions defined in [3, p. 1053], with which [3, Thm. 8.1] is concerned, because of the arbitrary σ in (4.2); the same proof of [3, Thm. 8.1] yields this result holding for $f(z)\in H(A;C)$, $A\geq 0$, also.)

If C is an open convex cone such that [10, property (C), p. 395] is satisfied by each compact subcone of C, then $f(x+iy)\in H(0;C)$ obtains a unique weak distributional boundary value in $(\mathcal{S}^\alpha)'$ as $y\to 0$, $y\in C$, for $\alpha=(\alpha_1,\cdots,\alpha_n)$ such that $\alpha_j\geq 1, j=1,\cdots,n$, by Theorem 4.5 in §4 of this paper. Similarly if [26, property P, p. 235] is satisfied by each compact subcone of C then $f(x+iy)\in H(0;C)$ obtains a unique weak distributional boundary value in $(W^\Omega)'$ as $y\to 0$, $y\in C$, by [26, Thm. 1].

We now state our results which correspond to Theorem 5.2 for the spaces $\mathcal{Z}'$, K'_r, $r\geq 1$, $(\mathcal{S}^\alpha)'$, and $(W^\Omega)'$.

THEOREM 6.1. *Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. Let $f(z)\in H(0;C)$. Let the unique weak $\mathcal{Z}'$ or K'_r, $r\geq 1$, (weak and strong in the case of K'_r) boundary value of $f(z)$, which exists, be $h(x)\in L^p$, $1\leq p\leq\infty$. Then $f(z)\in H^p(T^C)$, $1\leq p\leq\infty$, and*

$$f(z)=\int_{\mathbf{R}^n} h(t)Q(z;t)\,dt, \qquad z\in T^C. \tag{6.1}$$

THEOREM 6.2. *Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line and such that every compact subcone of C satisfies* [10, property (C), p. 395] (or [26, property P, p. 235].) *Let $f(z)\in H(0;C)$. Let the unique weak $(\mathcal{S}^\alpha)'$ (or $(W^\Omega)'$) boundary value of $f(z)$, which exists, be $h(x)\in L^p$, $1\leq p\leq\infty$. Then $f(z)\in H^p(T^C)$, $1\leq p\leq\infty$, and* (6.1) *holds.*

As noted in [10, p. 395] the first quadrant, and in fact any quadrant C_μ, and the forward (backward) light cone are examples of cones which satisfy the hypotheses in Theorem 6.2 corresponding to $(\mathcal{S}^\alpha)'$. Similarly these cones also satisfy the hypotheses with respect to $(W^\Omega)'$.

The proofs of Theorems 6.1 and 6.2 are obtained in the same manner that Theorem 5.2 was proved. We first prove Theorems 6.1 and 6.2 for C being an open convex cone which is contained in or is any of the 2^n quadrants C_μ in $\mathbf{R}^n$ as in Theorem 5.1. In this case we construct $g_\varepsilon(z)=f(z)/X_\varepsilon(z)$ as in (5.2) and $G_\varepsilon(t)$ as in (5.6). Using the growth (4.2) of $f(z)\in H(0;C)$, the same properties on $G_\varepsilon(t)$ as in the proof of Theorem 5.1 follow; and as the reader probably suspects, the proofs are completed in the same way as the proof of Theorem 5.1 is obtained. The only difference here is that some technical points with respect to the various distribution topologies involved need to be checked which correspond to facts already obtained in the proof of Theorem 5.1 We have verified all new points which are needed for the proofs of Theorems 6.1 and 6.2, and these become obvious to the reader as he reads through the proof of Theorem 5.1. Since the proofs of Theorems 6.1 and 6.2 are obtained in the same way as that of

Theorem 5.1 for the case that $C \subseteq C_\mu$, we ask the reader to reread the proof of Theorem 5.1 and verify any details desired for Theorems 6.1 and 6.2. Then, of course, Theorems 6.1 and 6.2 for arbitrary C as hypothesized follow in exactly the same way that Theorem 5.2 followed from Theorem 5.1. We do note that the recovery step (5.31) for $(f(z)-G(z))$ for the case that $C \subseteq C_\mu$ in Theorems 6.1 and 6.2 follows by Corollary 4.1 of this paper for the case $\mathcal{Z}'$, by [3, Thm. 8.2] for the case K_r' where we note that [3, (8.10), p. 1063] actually holds for all $z \in T^C$ since C is open, by Theorem 4.6 in this paper for the case $(\mathcal{S}^\alpha)'$, and by [26, Thm. 2] for the case $(W^\Omega)'$ where [26, Thm. 2, (ii)] actually holds for all $z \in T^C$ since C is open and $m > 0$ is arbitrary there.

It is obvious that corollaries to Theorems 6.1 and 6.2 can now be stated like Corollaries 5.1 and 5.2. As a converse to Theorems 6.1 and 6.2 we state the following result whose proof is obtained by exactly the same means as indicated for the proof of Theorem 5.3.

THEOREM 6.3. *Let C be an open convex cone such that $\overline{C}$ does not contain any entire straight line. Let $f(z) \in H^p(T^C)$, $1 \le p \le \infty$. There exists $h(x) \in L^p$ such that* (5.44) *holds in L^p if $1 \le p < \infty$ and in the weak-star topology of L^∞ if $p = \infty$. For all p, $1 \le p \le \infty$,* (5.44) *holds in the weak topology of $\mathcal{Z}'$, $(\mathcal{S}^\alpha)'$, and $(W^\Omega)'$ and in the weak and strong topology of K_r', $r \ge 1$. Further, if $1 \le p < \infty$ then $f(z)$ satisfies* (3.17), (4.7), *and* (4.2) *with $A = 0$; if $p = \infty$, $f(z)$ is bounded on T^C; and for all p, $1 \le p \le \infty$,* (6.1) *holds.*

Of course [28, Thm. 2] is a special case of our analysis in this section, and our results are applicable to the paper [13] where the distribution spaces $\mathcal{S}_1'$ and $(\mathcal{S}^1)'$ are considered.

REFERENCES

[1] E. J. BELTRAMI AND M. R. WOHLERS, *Distributional boundary value theorems and Hilbert transforms*, Arch. Rational Mech. Anal., 18 (1965), pp. 304–309.

[2] S. BOCHNER AND K. CHANDRASEKHARAN, *Fourier Transforms*, Princeton University Press, Princeton, NJ, 1949.

[3] R. D. CARMICHAEL, *Analytic functions related to the distributions of exponential growth*, this Journal, 10 (1979), pp. 1041–1068.

[4] ———, *Generalized Cauchy and Poisson integrals and distributional boundary values*, this Journal, 4 (1973), pp. 198–219.

[5] ———, *Functions analytic in an octant and boundary values of distributions*, J. Math. Anal. Appl., 33 (1971), pp. 616–626.

[6] ———, *Distributional boundary values of functions analytic in tubular radial domains*, Indiana Univ. Math. J., 20 (1971), pp. 843–853.

[7] ———, *Distributions of exponential growth and their Fourier transforms*, Duke Math. J., 40 (1973), pp. 765–783.

[8] R. D. CARMICHAEL AND E. K. HAYASHI, *A pointwise growth estimate for analytic functions in tubes*, Internat. J. Math. Math. Sci., 3 (1980), pp. 575–581.

[9] ———, *Analytic functions in tubes which are representable by Fourier–Laplace integrals*, Pacific J. Math., 90 (1980), pp. 51–61.

[10] R. D. CARMICHAEL AND E. O. MILTON, *Distributional boundary values in the dual spaces of spaces of type* $\mathcal{S}$, Pacific J. Math., 56 (1975), pp. 385–422.

[11] R. D. CARMICHAEL AND S. P. RICHTERS, *Growth of H^p functions in tubes*, Internat. J. Math. Math. Sci., 4 (1981), pp. 435–443.

[12] R. D. CARMICHAEL AND W. W. WALKER, *Representation of distributions with compact support*, Manuscripta Math., 11 (1974), pp. 305–338.

[13] F. CONSTANTINESCU, *Analytic properties of nonstrictly localizable fields*, J. Math. Phys., 12 (1971), pp. 293–298.

[14] R. E. EDWARDS, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.

[15] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[16] A. FRIEDMAN, *Generalized Functions and Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
[17] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions*, Vol. 1, Academic Press, New York, 1964.
[18] _____, *Generalized Functions*, Vol. 2, Academic Press, New York, 1968.
[19] _____, *Generalized Functions*, Vol. 3, Academic Press, New York, 1967.
[20] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[21] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, John Wiley and Sons, New York, 1968.
[22] A. KORÁNYI, *A Poisson integral for homogeneous wedge domains*, J. Analyse Math., 14 (1965), pp. 275–284.
[23] W. R. MADYCH, *Distributions with strong maximal functions in* $L^p(R^n)$, preprint.
[24] R. MEISE, *Darstellung temperierter vektorwertiger Distributionen durch holomorphe Funktionen* I, Math. Ann., 198 (1972), pp. 147–159.
[25] _____, *Darstellung temperierter vektorwertiger Distributionen durch holomorphe Funktionen* II, Math. Ann., 198 (1972), pp. 161–178.
[26] R. S. PATHAK, *Analytic functions having distributional boundary values in W'-spaces*, Math. Proc. Cambridge Philos. Soc., 87 (1980), pp. 227–242.
[27] B. E. PETERSEN, *On the Laplace transform of a temperate distribution supported by a cone*, Proc. Amer. Math. Soc., 35 (1972), pp. 123–128.
[28] A. K. RAINA, *On the role of Hardy spaces in form factor bounds*, Lett. Math. Phys., 2 (1978), pp. 513–519.
[29] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics, Vol.* II: *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.
[30] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.
[31] _____, *Mathematics for the Physical Sciences*, Addison-Wesley, Reading, MA, 1966.
[32] B. SIMON, *The* $P(\phi)_2$ *Euclidean (Quantum) Field Theory*, Princeton University Press, Princeton, NJ, 1974.
[33] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[34] R. F. STREATER AND A. S. WIGHTMAN, *PCT, Spin and Statistics, and All That*, W. A. Benjamin, New York, 1964.
[35] H. G. TILLMANN, *Darstellung der Schwartzschen Distributionen durch analytische Funktionen*, Math. Z., 77 (1961), pp. 106–124.
[36] V. S. VLADIMIROV, *Methods of the Theory of Functions of Many Complex Variables*, MIT Press, Cambridge, MA, 1966.

SIAM J. MATH. ANAL.
Vol. 14, No. 3, May 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1403-0014 $01.25/0

NECESSARY AND SUFFICIENT CONDITIONS RELATING THE COEFFICIENTS IN THE RECURRENCE FORMULA TO THE SPECTRAL FUNCTION FOR ORTHOGONAL POLYNOMIALS*

JEFFREY S. GERONIMO† AND PAUL G. NEVAI‡

Abstract. The problem considered here is, given the coefficients in the recurrence formula for polynomials orthogonal on a segment of the real line, what can be said about the spectral function with respect to which they are orthogonal? The coefficients are assumed to converge at a particular rate and the consequences for the spectral function are found that are necessary and sufficient.

AMS (MOS) subject classification (1970). Primary 42A52.

1. Introduction. Let $\rho(\lambda)$ be a nondecreasing function on a segment $[a,b]$ of the real line with infinitely many points of increase. Furthermore, let

$$s_n=\int_a^b \lambda^n d\rho(\lambda) \tag{1.1}$$

exist for all n. It is well known [5] that one can construct a unique set of polynomials $\{p(\lambda,n)\}$ with the following properties:

A) $p(\lambda,n)$ is a polynomial of degree n with positive leading coefficient,

B) $\int_a^b p(\lambda,n)p(\lambda,m)d\rho(\lambda)=\delta_{n,m}, \qquad n=0,1,2\cdots.$

Furthermore, these polynomials satisfy the following three-term recurrence formula:

(1.2)

$$a(n+1)p(\lambda,n+1)+b(n)p(\lambda,n)+a(n)p(\lambda,n-1)=\lambda p(\lambda,n), \qquad n=0,1,\cdots,$$

$$p(\lambda,0)=K(0)=\frac{1}{\sqrt{s_0}}, \qquad p(\lambda,-1)=0.$$

Here

$$a(n)=\int_a^b \lambda p(\lambda,n)p(\lambda,n-1)\,d\rho(\lambda) \tag{1.3}$$

and

$$b(n)=\int_a^b \lambda p(\lambda,n)^2 d\rho(\lambda). \tag{1.4}$$

In this paper we assume that the coefficients in the recurrence formula converge at a particular rate, and we find conditions for the spectral function that are necessary and sufficient. More precisely, let

$$\nu(0)=1, \qquad \nu(n)\geq 1 \tag{1.5}$$

* Received by the editors November 6, 1978, and in final revised form January 16, 1981.

† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The research of this author was supported by the National Science Foundation under grant MCS 81-01720.

be an even funciton of n with the following properties

$$\nu(n)\leq\begin{cases}\nu(n+1), & n\geq 0,\\ \nu(m)\nu(n-m), & n,m\geq 0,\end{cases} \tag{1.6}$$

and

$$\limsup_{n\to\infty}(\nu(n))^{1/n}=1.$$

We prove the following:

THEOREM 1. *Let* $0<K(0)<\infty$. *Let* $\{a(n)\}_1^\infty$ *be a sequence of positive numbers and* $\{b(n)\}_0^\infty$ *be a sequence of real numbers satisfying the following conditions*:

$$\lim_{n\to\infty}a(n)=a(\infty)>0, \qquad \lim_{n\to\infty}b(n)=b(\infty), \tag{1.7}$$

and

$$\sum_{n=1}^{\infty}n\nu(2n)\left\{\left|1-\frac{a(n)^2}{a(\infty)^2}\right|+|B(n-1)|\right\}<\infty. \tag{1.8}$$

Then a necessary and sufficient condition for the above to hold is that there exists a bounded nondecreasing function $\rho(\lambda)$ *on a finite segment of the real line of the form*

$$d\rho(\lambda)=\begin{cases}\sigma(\theta)\,d\lambda, & \lambda=2a(\infty)\cos\theta+b(\infty),\quad 0\leq\theta\leq\pi,\\ \sum_{i=1}^{N}\rho_i\delta(\lambda-\lambda_i)\,d\lambda, & \begin{cases}\lambda \text{ not as above},\\ \lambda_i\neq\lambda_j,\quad i\neq j,\quad \rho_i>0,\\ N<\infty,\end{cases}\end{cases} \tag{1.9}$$

where

$$\frac{\sigma(\theta)|d(\theta)|^2}{\sin\theta}=\frac{\sigma(-\theta)|d(-\theta)|^2}{\sin(-\theta)},$$

$\ln(\sigma(\theta)|d(\theta)|^2/\sin\theta)$ *has an absolutely convergent Fourier series, and*

$$\sum_{n=-\infty}^{\infty}|n|\nu(n)|q(n)-q(n+2)|<\infty. \tag{1.10}$$

The $a(n)$*'s and* $b(n)$*'s are related to* $\rho(\lambda)$ *by* (1.3) *and* (1.4).
Here

$$\frac{\sigma(\theta)|d(\theta)|^2}{\sin\theta}=\sum_{n=-\infty}^{\infty}q(n)e^{in\theta}, \tag{1.11}$$

$$B(n)=\frac{b(n)-b(\infty)}{a(\infty)}, \tag{1.12}$$

$\delta(\lambda)$ is the Dirac delta function, and $d(\theta)$ is equal to

1) 1 if $\sigma(\theta)$ is bounded at $\theta=0$ and $\theta=\pi$,
2) $1-e^{i\theta}$ if $\sigma(\theta)$ is unbounded at $\theta=0$,
3) $1+e^{i\theta}$ if $\sigma(\theta)$ is unbounded at $\theta=\pi$,
4) $1-e^{2i\theta}$ if $\sigma(\theta)$ is unbounded at $\theta=0$, and $\theta=\pi$.

In [3], (1.7) and (1.8) were assumed to hold with $\nu(n)=1$ for all n. The spectral function was constructed and shown to satisfy all the necessary conditions except

(1.10). Therefore, in §2 we briefly review the results of [3] and develop the equations necessary to complete the proof. Next (§3), we show that if one forms a new spectral function $\rho^0(\lambda)$ by adding or substracting a finite number of mass points to $d\rho(\lambda)$, then the coefficients $a^0(n)$, $b^0(n)$ associated with $\rho^0(\lambda)$ still satisfy (1.7) and (1.8). Finally, in §4 we show that one can alter the absolutely continuous part of $\rho(\lambda)$, $\sigma(\theta)$, in a way that allows one to call upon [2, Thm. 1] to complete the proof (§5).

2. A review. It is easy to see that if (1.7) holds, then the orthogonal polynomials $\{p(\lambda,n)\}$ satisfy the following recurrence formula:

$$\Phi(Z,n)=C(n)\Phi(Z,n-1), \tag{2.1}$$

where

$$C(n)=\frac{a(\infty)}{a(n)}\begin{bmatrix} Z-B(n-1) & 1/Z \\ \left\{\left(1-\dfrac{a(n)^2}{a(\infty)^2}\right)Z-B(n-1)\right\} & 1/Z \end{bmatrix}, \qquad n=1,2,\cdots, \tag{2.2}$$

$$\Phi(Z,n)=\begin{pmatrix} p(\lambda,n) \\ \psi(Z,n) \end{pmatrix}, \tag{2.3}$$

and

$$\lambda=a(\infty)\left(Z+\frac{1}{Z}\right)+b(\infty).$$

Z is a complex number. For initial conditions one takes

$$p(\lambda,0)=\psi(Z,0)=K(0)>0, \tag{2.4}$$

Two other useful solutions [3] of (2.1) are

$$\Phi_+(Z,n)=\begin{pmatrix} p_+(Z,n) \\ \psi_+(Z,n) \end{pmatrix} \tag{2.5}$$

and

$$\Phi_-(Z,n)=\begin{pmatrix} p_-(Z,n) \\ \psi_-(Z,n) \end{pmatrix}, \tag{2.6}$$

satisfying the following boundary conditions:

$$\begin{aligned} &\lim_{n\to\infty}|p_\pm(Z,n)-Z^{\pm n}|=0, && |Z|\leqq 1, \\ &\lim_{n\to\infty}|\psi_+(Z,n)|=0, && |Z|\leq 1, \\ &\lim_{n\to\infty}|\psi_-(Z,n)-(1-Z^2)Z^{-n}|=0, && |Z|\geq 1. \end{aligned} \tag{2.7}$$

It can be shown [3] that $\Phi_+(Z,n)$ and $\Phi_-(Z,n)$ are linearly independent for $Z\neq\pm 1$ and that

$$\Phi(Z,n)=\frac{K(0)}{a(\infty)(Z-1/Z)}[f_-(Z)\Phi_+(Z,n)-f_+(Z)\Phi_-(Z,n)], \qquad |Z|=1, \tag{2.8}$$

where

$$f_+(Z)=\frac{a(\infty)}{K(0)Z}\lim_{n\to\infty} Z^n\psi(Z,n), \tag{2.9}$$

and

$$f_-(Z)=\overline{f_+(Z)}=f_+\left(\frac{1}{Z}\right), \qquad |Z|=1. \tag{2.10}$$

One can include polynomials of the second kind in this scheme by defining

$$\Phi_\alpha(Z,n)=\begin{pmatrix} Q(Z,n) \\ \psi_\alpha(Z,n)\end{pmatrix}, \qquad n\geq 1, \tag{2.11}$$

satisfying (2.1) with boundary conditions

$$Q(\lambda,1)=\psi_\alpha(Z,1)=\frac{1}{a(1)}K(0). \tag{2.12}$$

It is now possible to write (see [3, App. B])

$$p_+(Z,n)=K(0)\left[f_{+\alpha}(Z)p(\lambda,n)-f_+(Z)Q(\lambda,n)\right], \qquad n\geq 1, \quad |Z|=1, \tag{2.13}$$

where

$$f_{+\alpha}(Z)=\frac{a(\infty)}{K(0)Z}\lim_{n\to\infty} Z^n\psi_\alpha(Z,n). \tag{2.14}$$

To proceed further it is convenient at this point to introduce the techniques of Banach algebras. Let A_ν denote the class of function integrable on $-\pi\leq\theta\leq\pi$ such that if g is an element of A_ν then

$$g(\theta)\approx\sum_{K'=-\infty}^{\infty} g(K)e^{iK\theta} \tag{2.15}$$

with

$$\|g\|_\nu=\sum_{K=\equiv\infty}^{\infty}\nu(K)|g(K)|<\infty. \tag{2.16}$$

$\nu(n)$ is defined in (1.5) and (1.6). Let A_ν^+ and A_ν^- denote those functions in A_ν of the form

$$g(\theta)\approx\sum_{K=0}^{\infty} g(K)e^{iK\theta} \tag{2.17}$$

and

$$h(\theta)\approx\sum_{K=-\infty}^{0} h(K)e^{iK\theta} \tag{2.18}$$

respectively.

Let $\|g\|_\nu$ be the norm on A_ν, A_ν^+ and A_ν^-. Then A_ν, A_ν^+ and A_ν^- are Banach algebras [1]. A will denote the Banach algebra where $\nu(n)=1$ for all n.

Returning now to (2.1), we can show the following [3]:

THEOREM 2. *If* (1.7) *holds and*

$$\sum_{n=1}^{\infty} n\nu(2n)\left\{\left|1-\frac{a(n)^2}{a(\infty)^2}\right|+|B(n-1)|\right\}<\infty, \tag{2.19}$$

then[1]

$$\lim_{n\to\infty}\left\|\frac{a(\infty)}{K(0)}Z^n\psi(Z,n)-Zf_+(Z)\right\|_\nu=0, \tag{2.20}$$

(2.21) $Zf_+(Z)$ *is analytic inside and continuous on the unit circle,*

(2.22) $p_+(e^{i\theta},n)\in A^+$ *for all* $n\geq 0$.

(2.23) *If* $f_+(Z)$ *has zeros inside the unit circle, they are,*
a) *real,*
b) *simple,*
c) *finite in number.*

Finally, if $f_+(Z)$ *has zeros on the unit circle,*
a) *they must be at* $Z=+1$ *and/or* $Z=-1$,
b) $Zf_+(Z)/d(Z)\in A^+$, $Z=e^{i\theta}$.

$d(Z)$ is defined following (1.12). The consequences for the spectral function, $\rho(\lambda)$, of the above are as follows:

THEOREM 3. *If* (1.7) *and* (2.19) *hold then there exists a bounded nondecreasing spectral function* $\rho(\lambda)$ *on a finite segment of the real line of the form given by* (1.9) *with*

$$\sigma(\lambda)=\frac{a(\infty)\sin\theta}{K(0)^2\pi|f_+(Z)|^2},\qquad \lambda=a(\infty)(e^{i\theta}+e^{-i\theta})+b(\infty),\quad Z=e^{i\theta} \tag{2.24}$$

and

$$\ln\left(\frac{\sigma(\theta)}{\sin\theta}|d(\theta)|^2\right)\in A. \tag{2.25}$$

3. The mass points. Let

$$d\rho^i(\lambda)=\begin{cases}\sigma(\lambda)\,d\lambda, & a\leq\lambda\leq b,\\ \sum_{m=1}^{N+i}\rho_m\delta(\lambda-\lambda_m)\,d\lambda, & \begin{cases}\lambda \text{ not as above},\\ i=-1,0,1,\\ N\geq 1,\end{cases}\end{cases} \tag{3.1}$$

with

$$|\lambda_m|\geq|\lambda_{m-1}|. \tag{3.2}$$

Furthermore, let $\{a_i(n)\}$ and $\{B_i(n)\}$ be the coefficients in (2.1) associated with $d\rho^i(\lambda)$ respectively.

THEOREM 4. *If*

$$\sum_{n=1}^{\infty}n\nu(2n)\left\{\left|1-\frac{a_0(n)^2}{a(\infty)^2}\right|+|B_0(n-1)|\right\}<\infty,$$

then

$$\sum_{n=1}^{\infty}n\nu(2n)\left\{\left|1-\frac{a_{-1}(n)^2}{a(\infty)^2}\right|+|B_{-1}(n-1)|\right\}<\infty.$$

[1] This requires a minor modification of the techniques used in [3, App. B].

Proof. Let

$$P_i(\lambda,n)=K_i(n)\lambda^n+\cdots, \qquad i=-1,0,+1, \tag{3.3}$$

be the polynomials orthonormal with respect to $d\rho^i(\lambda)$. One finds (see [4, §7])

$$p_{-1}(\lambda,n)=\frac{K_0(n)}{K_{-1}(n)}\left[p_0(\lambda,n)+\frac{\rho_N p_0(\lambda_N,n)K_n^0(\lambda,\lambda_N)}{1-\rho_N K_n^0(\lambda_N,\lambda_N)}\right], \tag{3.4}$$

with

$$K_n^0(\lambda,\lambda')=\sum_{j=0}^{n} p_0(\lambda,j)p_0(\lambda',j). \tag{3.5}$$

Solving for $\psi_i(Z,n)$ in the upper component of (2.1) then substituting it into the lower component yields

$$\psi_i(Z,n)=p_i(\lambda,n)-\frac{a_i(n)}{a(\infty)}Zp_i(\lambda,n-1). \tag{3.6}$$

(Note that $a_i(\infty)=a(\infty)$ since $d\rho^i(\lambda)$ all have the same absolutely continuous part [3].) It follows from (3.4) and (3.6) that

(3.7)

$$\psi_{-1}(Z,n)=\frac{K_0(n)}{K_{-1}(n)}\left[p_0(\lambda,n)+\frac{\rho_N p_0(\lambda_N,n)K_n^0(\lambda,\lambda_n)}{1-\rho_N K_n^0(\lambda_N,\lambda_N)}\right]$$
$$-\frac{a_{-1}(n)}{a(\infty)}\frac{K_0(n-1)}{K_{-1}(n-1)}Z\left[p_0(\lambda,n-1)+\frac{\rho_N p_0(\lambda_N,n-1)K_{n-1}^0(\lambda,\lambda_N)}{1-\rho_N K_{n-1}^0(\lambda_N,\lambda_N)}\right].$$

Since

$$a_i(n)=\frac{K_i(n-1)}{K_i(n)}, \tag{3.8}$$

(3.7) can be arranged to equal

(3.9)

$$K_{-1}(n)\psi_{-1}(Z,n)$$
$$=K_0(n)\left[\psi_0(Z,n)+\rho_N\left(\frac{p(\lambda_N,n)K_n^0(\lambda,\lambda_N)}{1-\rho_N K_n(\lambda_N,\lambda_N)}-\frac{a_0(n)}{a(\infty)}\frac{Zp(\lambda_N,n-1)K_{n-1}^0(\lambda,\lambda_N)}{1-\rho_N K_{n-1}^0(\lambda_N,\lambda_N)}\right)\right].$$

Multiplying by Z^n then equating coefficients of Z^{2n} gives

(3.10)

$$K_{-1}(n)K_{-1}(2n,2n)$$
$$=K_0(n)K_0(2n,2n)$$
$$+K_0(n)\hat{K}_0(2n,2n)\rho_N\left[\frac{p_0(\lambda_N,n)^2}{1-\rho_N K_n^0(\lambda_N,\lambda_N)}-\frac{a_0(n)^2}{a(\infty)^2}\frac{p_0(\lambda_N,n-1)^2}{1-\rho_N K_{n-1}^0(\lambda_N,\lambda_N)}\right],$$

where

$$Z^n\psi_i(Z,n)=\sum_{j=0}^{2n} K_i(2n,j)Z^j, \qquad i=-1,0,1, \tag{3.11}$$

and

$$Z^n p_i(\lambda,n)=\sum_{j=0}^{2n}\hat{K}_i(2n,j)Z^j,\qquad i=-1,0,1.$$

At λ_N, [3],

$$p_0(\lambda_N,n)=\frac{K_0(0)}{f^0_{+\alpha}}p^0_+(Z_N,n), \tag{3.12}$$

and

$$1-\rho_N K^0_\infty(\lambda_N,\lambda_N)=0. \tag{3.13}$$

Therefore the term in brackets in (3.10) equals

$$\left[\frac{p^0_+(Z_N,n)^2}{\Sigma_{j=n+1}^\infty p_+(Z_N,j)^2}-\frac{a_0(n)^2}{a(\infty)^2}\frac{p^0_+(Z_N,n-1)^2}{\Sigma_{j=n}^\infty p^0_+(Z_N,j)^2}\right]. \tag{3.14}$$

Solving (3.6) for $p^0_+(Z_N,n-1)$ then substituting into (3.14) gives

$$\begin{aligned}&=p^0_+(Z_N,n)^2\left\{\frac{\Sigma_{i=n}^\infty\left(p^0_+(Z_N,i)^2-\frac{1}{Z_N^2}p^0_+(Z_N,i+1)^2\right)}{\Sigma_{i=n+1}^\infty\Sigma_{j=n}^\infty p^0_+(Z_N,i)^2p^0_+(Z_N,j)^2}\right\}\\&\quad-\frac{2p^0_+(Z_N,n)\psi^0_+(Z_N,n)}{Z_N^2\Sigma_{i=n}^\infty p^0_+(Z_N,i)^2}+\frac{\psi^0_+(Z_N,n)^2}{Z_N^2\Sigma_{i=n}^\infty p^0_+(Z_N,i)^2}.\end{aligned} \tag{3.15}$$

Using (3.6) once again gives

$$\begin{aligned}&=p^0_+(Z_N,n)^2\left\{\sum_{i=n}^\infty\left(1-\frac{a_0(i+1)^2}{a(\infty)^2}\right)p^0_+(Z_N,i)^2\right.\\&\quad\left.-2\sum_{i=n}^\infty\frac{p^0_+(Z_N,i+1)\psi^0_+(Z_N,i+1)}{Z_N^2}+\sum_{i=n}^\infty\frac{\psi^0_+(Z_N,i+1)^2}{Z_N^2}\right\}\\&\qquad\Big/\sum_{i=n+1}^\infty\sum_{j=n}^\infty p^0_+(Z_N,i)^2p^0_+(Z_N,j)^2\\&\quad-\frac{2p^0_+(Z_N,n)\psi^0_+(Z_N,n)}{Z_N^2\Sigma_{i=n}^\infty p^0_+(Z_N,i)^2}+\frac{\psi^0_+(Z_N,n)^2}{Z_N^2\Sigma_{i=n}^\infty p^0_+(Z_N,i)^2}.\end{aligned} \tag{3.16}$$

From (2.1) one finds

$$\psi^0_+(Z_N,n)=\alpha_0(n)\sum_{j=n}^\infty\left\{\left(1-\frac{a_0(j+1)^2}{a(\infty)^2}\right)Z_N-B_0(j)\right\}Z_N^{j-n+1}\frac{p^0_+(Z_N,j)}{\alpha_0(j)}, \tag{3.17}$$

where

$$\alpha_0(n)=K_0(0)\prod_{i=1}^n\frac{a(\infty)}{a_0(i)}. \tag{3.18}$$

Since $p_+^0(Z,n)\in A^+$, $Z=e^{i\theta}$ it follows from (2.7), (3.2) and (3.12) that (see [3, App. B])

$$C^*|Z_N^n|\le|p_+^0(Z_N,n)|\le C|Z_N^n|, \qquad |Z_N|<1. \tag{3.19}$$

Substituting the above results into (3.10) gives

$$\begin{aligned}
&\left|K_{-1}(n)\frac{K_{-1}(2n,2n)}{K_0(n)}\right| \\
&\quad\le|K_0(2n,2n)| \\
&\qquad +C|\hat{K}_0(2n,2n)|\Bigg[\sum_{i=n}^{\infty}\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right||Z_N|^{2i-2n} \\
&\qquad\qquad +3\sum_{i=n}^{\infty}\sum_{j=i+1}^{\infty}\left\{\left|1-\frac{a_0(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}|Z_N|^{2j-2n} \\
&\qquad\qquad +3\sum_{j=n}^{\infty}\left\{\left|1-\frac{a_0(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}|Z_N|^{2j-2n}\Bigg].
\end{aligned} \tag{3.20}$$

It follows from the recurrence formula (2.1) that

$$K_i(2n,2n)=\left(1-\frac{a_i(n)^2}{a(\infty)^2}\right)\hat{K}_i(2n,2n). \tag{3.21}$$

Multiplying (3.4) by Z^n, then equating coefficients of Z^{2n} shows that $((K_{-1}(n)/K_0(n))\hat{K}_{-1}(2n,2n))$ is bounded from below. Therefore

$$\begin{aligned}
&\sum_{n=1}^{\infty}n\nu(2n)\left|1-\frac{a_{-1}(n)^2}{a(\infty)^2}\right| \\
&\quad\le C^+\Bigg[\sum_{n=1}^{\infty}n\nu(2n)\left|1-\frac{a_0(n)^2}{a(\infty)^2}\right| \\
&\qquad +3\sum_{j=2}^{\infty}j\nu(2j)\left\{\left|1-\frac{a_0(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}\sum_{n=1}^{j}(j-n)|Z_n|^{2j-2n} \\
&\qquad +3\sum_{j=1}^{\infty}j\nu(2j)\left\{\left|1-\frac{a_0(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}\sum_{n=1}^{j}|Z_N|^{2j-2n}\Bigg]<\infty.
\end{aligned} \tag{3.22}$$

To show that

$$\sum_{n=1}^{\infty}n\nu(2n)|B_{-1}(n-1)|<\infty, \tag{3.23}$$

multiply (3.9) by Z^n and equate coefficients of Z^{2n-1}:

$$
\begin{aligned}
&K_{-1}(n)K_{-1}(2n,2n-1)\\
&\quad=K_0(n)\Bigg[K_0(2n,2n-1)\\
&\qquad+\rho_N\hat{K}_0(2n,2n-1)\left\{\frac{p_0(\lambda_N,n)^2}{1-\rho_N K_n^0(\lambda_N,\lambda_N)}-\frac{a_0(n)^2}{a(\infty)^2}\,\frac{p_0(\lambda_N,n-1)^2}{1-\rho_N K_{n-1}^0(\lambda_N,\lambda_N)}\right\}\\
&\qquad+\rho_N\hat{K}_0(2n-2,2n-2)\left\{\frac{B_0(n-1)p_0(\lambda_N,n-1)^2}{1-\rho_N K_{n-1}^0(\lambda_N,\lambda_N)}+\frac{p_0(\lambda_N,n)p_0(\lambda_N,n-1)}{1-\rho_N K_n(\lambda_N,\lambda_N)}\right\}\\
&\qquad+\frac{-a_0(n)a_0(n-1)}{a(\infty)^2}\left\{\frac{p_0(\lambda_N,n-1)p_0(\lambda_N,n-2)}{1-\rho_N K_{n-1}(\lambda_N,\lambda_N)}\right\}\Bigg].
\end{aligned}
\tag{3.24}
$$

The upper component of (2.1) has been used to arrive at (3.24). The last two terms in (3.24) can be recast using (3.12), (3.13) and (3.6) to read

$$
\begin{aligned}
&\frac{p_0(\lambda_N,n)p_0(\lambda_N,n-1)}{1-\rho_N K_n(\lambda_N,\lambda_N)}-\frac{a_0(n)a_0(n-1)}{a(\infty)^2}\,\frac{p_0(\lambda_N,n-1)p_0(\lambda_N,n-2)}{1-\rho_N K_{n-1}(\lambda_N,\lambda_N)}\\
&\quad=\frac{a(\infty)}{a_0(n)Z_N}\Bigg[\frac{p_+^0(Z_N,n)^2}{\sum_{i=n+1}^{\infty}p_+(Z_N,i)^2}-\frac{a_0(n)^2}{a(\infty)^2}\,\frac{p_+^0(Z_N,n-1)^2}{\sum_{i=n}^{\infty}p_+(Z_N,i)^2}\\
&\qquad-\frac{\psi_+(Z_N,n)p_+(Z_N,n)}{\sum_{i=n+1}^{\infty}p_+(Z_N,i)^2}+\frac{a_0(n)^2}{a(\infty)^2}\,\frac{\psi_+(Z_N,n-1)p_+(Z_N,n-1)}{\sum_{i=n}^{\infty}p_+(Z_N,i)^2}\Bigg].
\end{aligned}
\tag{3.25}
$$

It follows from the recurrence formula (2.1) that

$$
K_i(2n,2n-1)=\frac{a(\infty)}{a_i(n)}\left(1-\frac{a_i(n)^2}{a(\infty)^2}\right)\hat{K}(2n-2,2n-3)-B_i(n-1)\hat{K}(2n,2n).
\tag{3.26}
$$

Therefore (3.25), (3.26), (3.24) and the previous analysis give (3.23).

THEOREM 5. *If*

$$
\sum_{n=1}^{\infty}n\nu(2n)\left\{\left|1-\frac{a_0(n)^2}{a(\infty)^2}\right|+|B_0(n-1)|\right\}<\infty,
$$

then

$$
\sum_{n=1}^{\infty}n\nu(2n)\left\{\left|1-\frac{a_1(n)^2}{a(\infty)^2}\right|+|B_1(n-1)|\right\}<\infty.
$$

Proof. Letting $\rho_N \to -\rho_{n+1}$ (see [4, §7]) in (3.10) yields

$$
\begin{aligned}
&K_1(n)K_1(2n,2n)\\
&\quad = K_0(n)K_0(2n,2n)\\
&\quad\quad - K_0(n)\hat{K}_0(2n,2n)\rho_{N+1}\left[\frac{p_0(\lambda_{N+1},n)^2}{1+\rho_{N+1}K_n^0(\lambda_{N+1},\lambda_{N+1})}\right.\\
&\quad\quad\quad \left. - \frac{a_0(n)^2}{a(\infty)^2}\,\frac{p_0(\lambda_{N+1},n-1)^2}{1+\rho_{N+1}K_{n-1}^0(\lambda_{N+1},\lambda_{N+1})}\right].
\end{aligned}
\tag{3.27}
$$

Setting

$$
D=\bigl(1+\rho_{N+1}K_n^0(\lambda_{N+1},\lambda_{N+1})\bigr)\bigl(1+\rho_{N+1}K_{n-1}^0(\lambda_{N+1},\lambda_{N+1})\bigr)
\tag{3.28}
$$

and rearranging the above equations gives

$$
\begin{aligned}
&= K_0(n)K_0(2n,2n)\\
&\quad - \frac{K_0(n)\hat{K}_0(2n,2n)\rho_{N+1}}{D}\left[p_0(\lambda_{N+1},n)^2 - \frac{a_0(n)^2}{a(\infty)^2}p_0(\lambda_{N+1},n-1)^2\right]\\
&\quad - \frac{K_0(n)\hat{K}_0(2n,2n)\rho_{N+1}^2}{D}\left[p_0(\lambda_{N+1},n)^2K_{n-1}^0(\lambda_{N+1},\lambda_{N+1})\right.\\
&\quad\quad \left. - \frac{a_0(n)^2}{a(\infty)^2}p_0(\lambda_{N+1},n-1)^2K_n^0(\lambda_{N+1},\lambda_{N+1})\right].
\end{aligned}
\tag{3.29}
$$

Letting $Z \to 1/Z$ in (3.6) then substituting it twice into the third term of (3.29) yields

$$
\begin{aligned}
&\left[p_0(\lambda_{N+1},n)^2K_{n-1}^0(\lambda_{N+1},\lambda_{N+1}) - \frac{a_0(n)^2}{a(\infty)^2}p_0(\lambda_{N+1},n-1)^2K_n^0(\lambda_{N+1},\lambda_{N+1})\right]\\
&\quad = -Z_{N+1}^2\psi_0\left(\frac{1}{Z_{N+1}},n\right)^2K_n^0(\lambda_{N+1},\lambda_{N+1})\\
&\quad\quad + 2Z_{N+1}^2p_0(\lambda_{N+1},n)\psi_0\left(\frac{1}{Z_{N+1}},n\right)K_n^0(\lambda_{N+1},\lambda_{N+1})\\
&\quad\quad - \left[\sum_{i=0}^{n-1}\left\{\left(1-\frac{a_0(i+1)^2}{a(\infty)^2}\right)p_0(\lambda_{N+1},i)^2\right.\right.\\
&\quad\quad\quad - 2Z_{N+1}^2\psi_0\left(\frac{1}{Z_{N+1}},i+1\right)p_0(\lambda_{N+1},i)\\
&\quad\quad\quad \left.\left. - Z_{N+1}^2\psi_0\left(\frac{1}{Z_{N+1}},i+1\right)^2\right\} - Z_{N+1}^2K(0)^2\right]p_0(\lambda_{N+1},n)^2.
\end{aligned}
\tag{3.30}
$$

It is clear from the recurrence formula (Z_N is outside the support of $d\rho^0$) that

$$C\rho_{N+1}|Z_{N+1}|^{-2n}<\rho_{N+1}p_0(\lambda_{N+1},n)^2<\rho_{N+1}K_n^0(\lambda_{N+1},\lambda_{N+1}) \tag{3.31}$$
$$<1+\rho_{N+1}K_n^0(\lambda_{N+1},\lambda_{N+1}),$$

and it can be shown (see [3, App. B]) that

$$|p_0(\lambda_{N+1},n)|\le C'|Z_{N+1}|^{-n}, \qquad |Z_{N+1}|<1. \tag{3.32}$$

Therefore (3.29) reduces to

$$\le|K_0(n)K_0(2n,2n)|+3C\hat{}\,\frac{|K_0(n)\hat{K}_0(2n,2n)|}{\rho_{N+1}}|Z_{N+1}|^{2n}$$
$$+C\hat{}\,|K_0(n)\hat{K}_0(2n,2n)|\left[6\sum_{i=1}^{n}|Z_{N+1}|^{2n-i}\left|\psi\left(\frac{1}{Z_{N+1}},i\right)\right|\right. \tag{3.33}$$
$$\left.+\sum_{i=0}^{n-1}\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right||Z_{N+1}|^{2n-2i}\right].$$

Here the fact that $Z^n\psi_0(1/Z,n)$ is bounded for all n has been used (see below). It follows from (2.1) and (3.32) that

$$\left|\psi_0\left(\frac{1}{Z_{N+1}},i\right)\right|\le C\left[|Z_{N+1}|^i+\sum_{j=0}^{i-1}\left\{\left|1-\frac{a(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}|Z_{N+1}|^{i-2j+1}\right]. \tag{3.34}$$

Therefore (3.33) becomes

$$\le|K_0(n)K_0(2n,2n)|+3C\hat{}\,\frac{|K_0(n)\hat{K}_0(2n,2n)|}{\rho_{N+1}}|Z_{N+1}|^{2n}$$
$$+C\hat{}\,|K_0(n)\hat{K}_0(2n,2n)|\left[6n|Z_{N+1}|^{2n}+\sum_{i=0}^{n-1}\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right||Z_{N+1}|^{2n-2i}\right. \tag{3.35}$$
$$\left.+6\sum_{i=1}^{n}\sum_{j=0}^{i-1}\left\{\left|1-\frac{a_0(j+1)^2}{a(\infty)^2}\right|+|B_0(j)|\right\}|Z_{N+1}|^{2n-2j}\right].$$

Multiplying (3.35) by $n\nu(2n)$ and summing on n, let us examine the last two terms on the right-hand side. We see that

$$\sum_{n=1}^{\infty}n\nu(2n)\sum_{i=0}^{n-1}\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right||Z_{N+1}|^{2n-2i}$$
$$\le\sum_{n=1}^{\infty}(n-i)\nu(2(n-i-1))|Z_{N+1}|^{2n-2i}\sum_{i=0}^{n-1}\nu(2(i+1))\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right| \tag{3.36}$$
$$+\sum_{n=1}^{\infty}\nu(2(n-i-1))|Z_{N+1}|^{2n-2i}\sum_{i=0}^{n-1}i\nu(2(i+1))\left|1-\frac{a_0(i+1)^2}{a(\infty)^2}\right|<\infty$$

and

$$\sum_{n=1}^{\infty} n\nu(2n) \sum_{i=1}^{n} \sum_{j=0}^{i-1} \left\{ \left| 1 - \frac{a_0(j+1)^2}{a(\infty)^2} \right| + |B_0(j)| \right\} |Z_{N+1}|^{2n-2j}$$

$$\leq \sum_{n=1}^{\infty} (n-j)^2 \nu(2(n-j-1)) |Z_{N+1}|^{2n-2j}$$

$$\times \sum_{j=0}^{n-1} \nu(2(j+1)) \left\{ \left| 1 - \frac{a_0(j+1)^2}{a(\infty)^2} \right| + |B_0(j)| \right\}$$

$$+ \sum_{n=1}^{\infty} (n-j)\nu(2(n-j-1)) |Z_{N+1}|^{2n-2j}$$

$$\times \sum_{j=0}^{n-1} j\nu(2(j+1)) \left\{ \left| 1 - \frac{a_0(j+1)^2}{a(\infty)^2} \right| + |B_0(j)| \right\} < \infty.$$

These two equations plus (3.21) and (1.6) imply (the fact that $(K_1(n)/K_0(n))\hat{K}_1(2n, 2n)$ is bounded from below follows from (3.4) with $\rho_N = -\rho_{N+1}$) that

$$\sum_{n=1}^{\infty} n\nu(2n) \left| 1 - \frac{a_1(n)^2}{a(\infty)^2} \right| < \infty. \tag{3.37}$$

To show that

$$\sum_{n=1}^{\infty} n\nu(2n) |B_1(n-1)| < \infty, \tag{3.38}$$

let $\rho_N \to -\rho_{N+1}$ in (3.24), then use (3.6) and (3.26) and the procedures that led to (3.37) to yield the desired result.

4. Alteration of the absolutely continuous part. Without loss of generality let us assume that $a(\infty) = \frac{1}{2}$ and $b(\infty) = 0$ in (1.8). Furthermore, let $f_+(Z) \neq 0$ for $|Z| < 1$. This implies that

$$d\rho(\lambda) = \sigma(\lambda)\, d\lambda \qquad (-1 \leq \lambda \leq 1). \tag{4.1}$$

If $f_+(1) = 0$, define

$$\sigma^*(\lambda) = (1-\lambda)\sigma(\lambda) \qquad (-1 \leq \lambda \leq 1), \tag{4.2}$$

and if $f_+(1) \neq 0$, define

$$\hat{\sigma}(\lambda) = \frac{\sigma(\lambda)}{1-\lambda} \qquad (-1 \leq \lambda \leq 1). \tag{4.3}$$

Let

$$p^*(\lambda, n) = K^*(n)\lambda^n + \cdots, \tag{4.4}$$

and

$$\hat{p}(\lambda, n) = \hat{K}(n)\lambda^n + \cdots \tag{4.5}$$

be the orthogonal polynomials associated with $\sigma^*(\lambda)$ and $\hat{\sigma}(\lambda)$ respectively. Finally let $a^*(n)$ and $b^*(n)$, and $\hat{a}(n)$ and $\hat{b}(n)$ be the coefficients in (1.2) for $p^*(\lambda, n)$ and $\hat{p}(\lambda, n)$ respectively.

THEOREM 6. *If*

$$\sum_{n=1}^{\infty} n\nu(2n)\left[|1-2a(n)|+|b(n)|\right]<\infty \tag{4.6}$$

and if $f_+(1)=0$, *then*

$$\sum_{n=1}^{\infty} n\nu(2n)\left[|1-2a^*(n)|+|b^*(n)|\right]<\infty. \tag{4.7a}$$

If (4.6) *holds and* $f_+(1)\neq 0$, *then*

$$\sum_{n=1}^{\infty} n\nu(2n)\left[|1-2\hat{a}(n)|+|\hat{b}(n)|\right]<\infty. \tag{4.7b}$$

Proof. First assume that $f_+(1)=0$. Expanding $(1-\lambda)p^*(\lambda,n)$ in a Fourier series in $\{p(\lambda,i)\}$ and then using the orthogonality relations, one finds

$$(1-\lambda)p^*(\lambda,n)=\frac{K(n)}{K^*(n)}p(\lambda,n)-\frac{K^*(n)}{K(n+1)}p(\lambda,n+1). \tag{4.8}$$

In particular, for $\lambda=1$,

$$\frac{K(n)K(n+1)}{K^*(n)^2}=\frac{p(1,n+1)}{p(1,n)}. \tag{4.9}$$

Therefore,

$$a^*(n)^2=a(n)a(n+1)\,\frac{p(1,n+1)p(1,n-1)}{p(1,n)^2}, \tag{4.10}$$

If substituting (2.13) into (4.9) and noting that $f_+(1)=0$ gives

$$a^*(n)^2=a(n)a(n+1)\,\frac{p_+(1,n+1)p_+(1,n-1)}{p_+(1,n)^2}. \tag{4.11}$$

Squaring (4.8) and integrating it against $\sigma(\lambda)$ gives

$$1-b^*(n)=\int_{-1}^{1}(1-\lambda)^2p^*(\lambda,n)^2\sigma(\lambda)\,d\lambda=\frac{K(n)^2}{K^*(n)^2}+\frac{K^*(n)^2}{K(n+1)^2}, \tag{4.12}$$

where (1.4) has been used. It follows from (4.9) that

$$b^*(n)=1-a(n+1)\left[\frac{p_+(1,n+1)}{p_+(1,n)}+\frac{p_+(1,n)}{p_+(1,n+1)}\right]. \tag{4.13}$$

Let us now find the corresponding expressions for $\hat{a}(n)$ and $\hat{b}(n)$. From the Fourier series expansion one finds

$$\hat{p}(\lambda,n)=-\frac{K(n-1)}{\hat{K}(n)}p(\lambda,n-1)+\frac{\hat{K}(n)}{K(n)}p(\lambda,n). \tag{4.14}$$

Multiplying by $\hat{\sigma}(\lambda)$ and integrating yields

$$\frac{K(n-1)K(n)}{\hat{K}(n)^2}=\frac{\int_{-1}^{1}p(\lambda,n)\,\frac{\sigma(\lambda)}{1-\lambda}\,d\lambda}{\int_{-1}^{1}p(\lambda,n-1)\,\frac{\sigma(\lambda)}{1-\lambda}\,d\lambda}. \tag{4.15}$$

Therefore,

$$(4.16)\quad \hat{a}(n+1)^2=a(n)a(n+1)\frac{\int_{-1}^{1}p(\lambda,n+1)\frac{\sigma(\lambda)}{1-\lambda}d\lambda\int_{-1}^{1}p(\lambda,n-1)\frac{\sigma(\lambda)}{1-\lambda}d\lambda}{\left[\int_{-1}^{1}p(\lambda)\frac{\sigma(\lambda)}{1-\lambda}d\lambda\right]^2}.$$

Since $Q(\lambda,n)$ is a polynomial of the second kind,

$$(4.17)\qquad Q(\lambda,n)=\int_{-1}^{1}\frac{(p(\lambda,n)-p(t,n))\sigma(t)\,dt}{\lambda-t}.$$

Substituting the above equation into (2.13) and using the fact that $f_+(Z)\neq 0$ for $|Z|\leq 1$, we get

$$(4.18)\qquad \lim_{n\to\infty}p(\lambda,n)=\infty,\qquad |Z|<1,$$

and

$$(4.19)\qquad \lim_{n\to\infty}p_+(Z,n)=0,\qquad |Z|<1,$$

which give

$$(4.20)\qquad f_{+\alpha}=f_+(Z)\int_{-1}^{1}\frac{\sigma(t)}{\lambda-t}dt.$$

Equation (2.13) now becomes

$$(4.21)\qquad p_+(1,n)=K(0)f_+(1)\int_{-1}^{1}\frac{p(t,n)}{1-t}\sigma(t)\,dt.$$

Hence

$$(4.22)\qquad \hat{a}(n+1)^2=a(n)a(n+1)\frac{p_+(1,n+1)p_+(1,n-1)}{p_+(1,n)^2}.$$

Squaring (4.14), multiplying by $\sigma(\lambda)$ and integrating give

$$(4.23)\qquad \hat{b}(n+1)=1-a(n+1)\left[\frac{p_+(1,n+1)}{p_+(1,n)}+\frac{p_+(1,n)}{p_+(1,n+1)}\right],$$

where (4.15) has been used. It follows from (4.11), (4.13), (4.22) and (4.23) that (4.6) and (4.7) hold if

$$(4.24)\qquad \sum_{n=2}^{\infty}n\nu(2n)\left|\frac{p_+(1,n+1)p_+(1,n-1)}{p_+(1,n)^2}-1\right|<\infty,$$

and

$$(4.25)\qquad \sum_{n=2}^{\infty}n\nu(2n)\left|\frac{p_+(1,n+1)}{p_+(1,n)}+\frac{p_+(1,n)}{p_+(1,n+1)}-2\right|<\infty.$$

Set

$$(4.26)\qquad \delta(n)=p_+(1,n+1)-p_+(1,n).$$

Then

$$\frac{p_+(1,n+1)p_+(1,n-1)}{p_+(1,n)^2}-1=\frac{\delta(n)-\delta(n-1)}{p_+(1,n)}-\frac{\delta(n)\delta(n-1)}{p_+(1,n)^2} \tag{4.27}$$

and

$$\frac{p_+(1,n+1)}{p_+(1,n)}+\frac{p_+(1,n)}{p_+(1,n+1)}-2=\frac{\delta(n)^2}{p_+(1,n)p_+(1,n+1)}. \tag{4.28}$$

From (2.13),

$$\lim_{n\to\infty} p_+(1,n)=1. \tag{4.29}$$

Therefore (4.24) and (4.25) converge if

$$\sum_{n=2}^{\infty} n\nu(2n)|\delta(n)-\delta(n-1)|<\infty \tag{4.30}$$

and

$$\sum_{n=2}^{\infty} n\nu(2n)\delta(n)^2<\infty. \tag{4.31}$$

It follows from (1.2) that

$$\begin{aligned}\delta(n)-\delta(n-1)&=p_+(1,n+1)-2p_+(1,n)+p_+(1,n-1)\\ &=[1-2a(n+1)]\,p_+(1,n+1)\\ &\quad -2b(n)p_+(1,n)+[1-2a(n)]\,p_+(1,n-1).\end{aligned} \tag{4.32}$$

Thus (1.8) and (4.29) imply that (4.30) holds. Equation (4.30) and the properties of $\nu(n)$ imply

$$\sum_{n=2}^{\infty} |\delta(n)|<\infty. \tag{4.33}$$

Therefore,

$$\begin{aligned}\sum_{n=2}^{\infty} n\nu(2n)\delta(n)^2&\le \sum_{n=2}^{\infty}|\delta(n)|\sum_{i=n}^{\infty} i\nu(2i)|\delta(i)-\delta(i-1)|\\ &<\sum_{n=2}^{\infty}|\delta(n)|\sum_{i=2}^{\infty} i\nu(2i)|\delta(i)-\delta(i-1)|.\end{aligned} \tag{4.34}$$

It is clear that the above procedure leads to the same results if $1-\lambda$ is replaced by $1+\lambda$.

Some straightforward examples of Theorem 4 are Chebyshev polynomials of the first and second kind, and the Jacobi polynomials $P_n^{(1/2,-1/2)}$ and $P_n^{(-1/2,1/2)}$ (see [5]).

5. Conclusions. The proof of Theorem 1 is now a consequence of [2, Thm. 1][2] and the theorems in the previous sections. Starting with (1.7) and (1.8), one constructs the spectral function using Theorem 3. To show the absolutely continuous part $\sigma(\lambda)$ of

[2] At this point we wish to point out an error in the literature. [2, Thm. 3.1 and consequently Lemma 3.1] are incorrect. The sufficiency part of [2, Thm. 1] follows from the argument in Appendix A of that paper.

$\rho(\lambda)$ satisfies (1.10), one defines a new weight

$$d\rho^0(\lambda)=\sigma^0(\lambda)\,d\lambda \tag{5.1}$$

where

$$\sigma^0(\lambda)=\frac{\sigma(\theta)}{\sin\theta}|d(\theta)|^2. \tag{5.2}$$

It is a consequence of the above sections that the coefficients $a^0(n)$ and $b^0(n)$ associated with $\sigma^0(\lambda)$ satisfy (1.7) and (1.8). $\sigma^0(\lambda)$ also satisfies the conditions required in [2, Thm. 1],[2] therefore (1.10) follows. Sufficiency is proved by reversing the steps.

REFERENCES

[1] G. BAXTER, *A convergence equivalence related to polynomials orthogonal on the unit circle*, Trans. Amer. Math. Soc., 99 (1961), pp. 471–487.

[2] J. S. GERONIMO, *A relation between the coeffcients in the recurrence formula and the spectral function for orthogonal polynomials*, Trans. Amer. Math. Soc., 260 (1980), pp. 65–82.

[3] J. S. GERONIMO AND K. M. CASE, *Scattering theory and polynomials orthogonal on the real line*, Trans. Amer. Math. Soc., 258 (1980), pp. 467–494.

[4] PAUL G. NEVAI, *Orthogonal Polynomials*, Memoirs, vol. 213, American Mathematical Society, Providence, RI, 1979.

[5] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, New York, Vol. 23, 1939.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0001 $01.25/0

HOW AN INITIALLY STATIONARY INTERFACE BEGINS TO MOVE IN POROUS MEDIUM FLOW*

D. G. ARONSON†, L. A. CAFFARELLI‡ AND S. KAMIN§

Abstract. The interface we study is the boundary of the support of the density of a gas flowing in a homogeneous porous medium. It is known that for certain initial distributions the interface remains stationary for a positive time. We derive upper and lower estimates for this waiting time and give a condition which is sufficient to guarantee that the interface begins to move in a smooth manner. In some cases our estimates give the exact waiting time.

Introduction. Isentropic flow of an ideal gas in a one-dimensional homogeneous porous medium is described by the degenerate nonlinear parabolic equation

$$\frac{\partial u}{\partial t}=\frac{\partial^2}{\partial x^2}(u^m). \tag{0.1}$$

Here u represents the density of the gas and $m\in[2,+\infty)$ is a constant (see, for example, [1]). For many mathematical questions concerning equation (0.1) the restriction $m\geq 2$ is irrelevant. Most of the basic results for this equation hold with $m>1$ and we shall make this assumption. The most striking manifestation of the nonlinearity and degeneracy of equation (0.1) is the finite speed of propagation of disturbances from rest. Specifically, if a solution $u=u(x,t)$ of equation (0.1) is such that $u(\cdot,0)$ has bounded support in $\mathbb{R}$, then $u(\cdot,t)$ has bounded support in $\mathbb{R}$ for all $t>0$. In general, $\operatorname{supp} u(\cdot,t)$ expands as t increases. Indeed this expansion always occurs if one waits long enough and, once begun, never stops. However, for certain initial conditions $\operatorname{supp} u(\cdot,t)$ will remain unchanged for a positive time. This is the situation which we shall consider in this paper. Our main results are criteria involving only the initial datum $u(\cdot,0)$ which allow us to estimate and, in some cases, predict exactly when $\operatorname{supp} u(\cdot,t)$ begins to expand, and criteria which allow us to assert that the expansion occurs smoothly.

We shall consider the initial value problem for equation (0.1). That is, given a function $u_0:\mathbb{R}\to[0,+\infty)$ we seek a function $u:\mathbb{R}\times[0,T)\to[0,+\infty)$ for some $T\in\mathbb{R}^+$ which satisfies

$$\frac{\partial u}{\partial t}=\frac{\partial^2}{\partial x^2}(u^m) \quad \text{in } \mathbb{R}\times(0,T), \qquad u(\cdot,0)=u_0 \quad \text{in } \mathbb{R}. \tag{0.2}$$

If u_0 is not strictly positive then, in general, the initial value problem (0.2) does not have a solution in the classical sense. It is, however, solvable in a suitable generalized sense. Specifically, u is said to be a *weak solution* of problem (0.2) in $\mathbb{R}\times(0,T)$ if

(i) u is continuous and nonnegative in $\mathbb{R}\times[0,T)$,

(ii) $(u^m)_x$ exists in the sense of distributions and is bounded in $\mathbb{R}\times[0,T)$,

* Received by the editors October 9, 1981.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by the National Science Foundation under grant MCS 80-02392.

‡ Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The work of this author was supported in part by the National Science Foundation under grant MCS 79-00985.

§ Department of Mathematics, Tel Aviv University, Tel Aviv, Israel. The work of this author was supported in part by the Air Force Office of Scientific Research under grant 78-3602.

(iii) u satisfies the integral identity

$$\iint_{\mathbb{R}\times(0,T)} \left\{ \frac{\partial(u^m)}{\partial x}\frac{\partial\varphi}{\partial x} - u\frac{\partial\varphi}{\partial t} \right\} dx\,dt = \int_{\mathbb{R}} u_0(x)\varphi(x,0)\,dx$$

for all $\varphi \in C^1(\mathbb{R}\times[0,T])$ which vanish for $|x|$ large and t near T.

It is known [16] that problem (0.2) has a unique weak solution if, for example, $u_0^m \in \mathrm{Lip}(\mathbb{R})$.

In order to discuss the behavior of $\operatorname{supp} u$ it is convenient to replace the density u by the (normalized) pressure, $v = mu^{m-1}/(m-1)$. If u satisfies (0.1) then v satisfies

$$\frac{\partial v}{\partial t} = (m-1)v\frac{\partial^2 v}{\partial x^2} + \left(\frac{\partial v}{\partial x}\right)^2. \tag{0.3}$$

We shall call $v: \mathbb{R}\times[0,T)\to[0,+\infty)$ a weak solution of the initial value problem

$$\begin{aligned} &\frac{\partial v}{\partial t} = (m-1)v\frac{\partial^2 v}{\partial x^2} + \left(\frac{\partial v}{\partial x}\right)^2 && \text{in } \mathbb{R}\times(0,T), \\ &v(\cdot,0) = v_0 && \text{in } \mathbb{R} \end{aligned} \tag{0.4}$$

if $u = \{(m-1)v/m\}^{1/(m-1)}$ is a weak solution of problem (0.2) with $u_0 = \{(m-1)v_0/m\}^{1/(m-1)}$. Kalashnikov [11] has shown that the weak solution of (0.4) is unique in the class of functions v which satisfy $0 \le v(x,t) \le Ax^2 + B$ in $\mathbb{R}\times[0,T)$ for some positive constants A and B.

Let $v = v(x,t)$ be a weak solution of problem (0.4) where we assume that

$$v_0(x)\begin{cases} = 0 & \text{for } x \in [0,+\infty), \\ > 0 & \text{for all sufficiently large } x < 0. \end{cases}$$

Define

$$t^* = \sup\{t \in [0,T) : v(0,t) = 0\}.$$

Knerr [12] has shown that

$$v_0(x) \ge c(-x)^\gamma \text{ on } (-\delta,0) \text{ for some } c \in \mathbb{R}^+, \delta \in \mathbb{R}^+ \text{ and } \gamma \in (0,2)$$

implies that $t^* = 0$, while

$$v_0(x) \le cx^2 \text{ on } (-\delta,0) \text{ for some } c \in \mathbb{R}^+ \text{ and } \delta \in \mathbb{R}^+$$

implies that $t^* > 0$. On the other hand, Aronson [2] and Kalashnikov [12] have shown that there exists a nondecreasing Lipschitz continuous function $\zeta: [0,T)\to[0,+\infty)$ such that $\zeta(0) = 0$ and

$$v(x,t)\begin{cases} = 0 & \text{for } x \ge \zeta(t), \\ > 0 & \text{for all sufficiently large } x < \zeta(t) \end{cases}$$

for each $t \in [0,T)$. The curve $x = \zeta(t)$ is the right-hand interface, and we shall focus our attention on it exclusively. We are interested in estimating and evaluating t^*, and in the smoothness of $\zeta(t)$.

Aronson [2] has shown that $\lim_{x\uparrow\zeta(t)} v_x(x,t)$ exists and Knerr [13] has shown that $D^+\zeta(t) = -v_x(\zeta(t)-,t)$ for each $t \in (0,T)$ where D^+ denotes the right-hand derivative. In addition, Knerr has proved that once the interface begins to move it must continue to move. Caffarelli and Friedman [8] have gone even further. They have proved that $D^+\zeta(\tau) > 0$ for all $\tau > t^*$ and that $\zeta \in C^1(t^*,T)$. We shall give a new and somewhat simplified proof of this result, as well as conditions which guarantee that $\zeta \in C^1(0,T)$.

Before stating our results, it will be useful to recall two explicit solutions of (0.3) which will play an important role in our work. Let $t_m = 1/2(m+1)$. For arbitrary $\alpha \in \mathbb{R}^+$,

$$v^*(x,t) = \begin{cases} \dfrac{t_m x^2}{(t_m/\alpha)-t} & \text{in } (-\infty,0]\times[0,t_m/\alpha), \\ 0 & \text{in } \mathbb{R}^+\times[0,t_m/\alpha) \end{cases}$$

is the unique weak solution of problem (0.4) in $\mathbb{R}\times[0,t_m/\alpha)$ with

$$v_0(x) = \begin{cases} \alpha x^2 & \text{in } (-\infty,0], \\ 0 & \text{in } \mathbb{R}^+. \end{cases}$$

This solution is due to Barenblatt [7]. For uniqueness, see Kalashnikov [11]. Here $T = t^* = t_m/\alpha$. As we shall show, for certain solutions v of problem (0.4), $v \sim v^*$ in the neighborhood of a vertical interface. The other special solution of (0.3) is the piecewise linear function

$$L_\gamma(x,t) \equiv (\gamma^2 t - \gamma x)^+$$

for $\gamma \in \mathbb{R}^+$. In Lemma 2.1 we show that at each point where the interface is not vertical, the right-hand derivative, $D^+\zeta$, determines the asymptotic behavior of the solution v of (0.3) in a uniform neighborhood of the point. Specifically, $D^+\zeta(t_0) = \gamma > 0$ implies that $v \sim L_\gamma$ in a neighborhood of $(\zeta(t_0), t_0)$. In addition to making comparisons with these special solutions, we shall rely heavily on similarity transformations. If $v = v(x,t)$ is a solution of (0.3), then for all $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$ so is

$$\frac{\beta}{\alpha^2} v(\alpha x, \beta t).$$

This observation permits us to use the limiting values of solutions under various deformations of the (x,t)-plane.

Our first result is an estimate for t^* from above in terms of the local behavior of v_0 near $x = 0$, and from below in terms of its global behavior.

THEOREM A. *Let v be a solution of problem* (0.4) *with $v_0 \equiv 0$ in $\mathbb{R}^+$. If*

$$v_0(x) = \alpha x^2 + o(x^2) \quad \text{as } x \uparrow 0 \tag{0.5}$$

and

$$v_0(x) \le \beta x^2 \quad \text{in } \mathbb{R}^- \tag{0.6}$$

for some constants $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$, then

$$\frac{t_m}{\beta} \le t^* \le \frac{t_m}{\alpha}.$$

An immediate consequence of Theorem A is the following result.

COROLLARY A.1. *If v_0 satisfies* (0.5),

$$v_0(x) \le \alpha x^2 \quad \text{in } \mathbb{R}^- \tag{0.7}$$

and $v_0 \equiv 0$ in $\mathbb{R}^+$, then

$$t^* = \frac{t_m}{\alpha}.$$

A remarkable feature of Corollary A.1 is that the waiting time t^* is completely determined by the conditions (0.5) and (0.7). It is therefore, in some sense, independent of the "size" of v_0; for example, t^* does not depend on the measure of $\operatorname{supp} v_0$.

A growth condition such as (0.6) is necessary in order to ensure the existence of a solution of problem (0.4) (cf. [5]). However, it is by no means clear that a condition as stringent as (0.7) is needed in order for us to conclude that $t^* = t_m/\alpha$.

The remainder of our results concern the smoothness of the interface. We first give a new proof of the following theorem which was first proved by Caffarelli and Friedman in [8].

THEOREM B [8]. *If v is a solution of problem* (0.4) *with $v_0 \equiv 0$ in $\mathbb{R}^+$ and $v_0(x) > 0$ for all sufficiently large $x < 0$, then*

$$\zeta \in C^1(t^*, T).$$

Our final results give criteria which guarantee that $\zeta \in C^1(0, T)$ even though $t^* > 0$.

THEOREM C. *Let v be a solution of problem* (0.4) *where v_0 satisfies $v_0(x) \equiv 0$ in $\mathbb{R}^+$,* (0.5), *and* (0.6) *for some $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$. If*

$$t^* = \frac{t_m}{\alpha} \tag{0.8}$$

and v_{xx} is a nondecreasing function of x in $(-\delta, 0) \times (0, t_m/\alpha)$ for some $\delta \in \mathbb{R}^+$, then

$$\zeta \in C^1(0, T).$$

Note that, in view of Corollary A.1, if v_0 satisfies (0.7) instead of (0.6), then (0.8) is automatically satisfied. However, as noted above, we strongly doubt the necessity of a growth condition as stringent as (0.7). The condition that v_{xx} be monotone near the interface also occurs in diBenedetto's work [9] on the regularity of v_t. It is also probably much too strong. The following corollaries to Theorem C give somewhat more practical criteria for the smoothness of the interface.

COROLLARY C.1. *Let v and $\bar{v}$ be solutions of problem* (0.4) *with $v(x, 0) = v_0(x)$ and $\bar{v}(x, 0) = \bar{v}_0(x)$. Suppose that $\bar{v}$ satisfies all of the hypothesis of Theorem* C *for some value of $\alpha \in \mathbb{R}^+$ and that v satisfies $v_0 \equiv 0$ in $\mathbb{R}^+$,* (0.5), (0.6) *and* (0.8) *for the same value of α. If $v_0 \le \bar{v}_0$ in $\mathbb{R}$, then*

$$\zeta \in C^1(0, T).$$

Corollary C.1 is useful only when one can construct a suitable comparison function $\bar{v}$. This can be done fairly simply in the following special case.

COROLLARY C.2. *Suppose that v_0 satisfies $v_0 \equiv 0$ in $\mathbb{R}^+$,* (0.5), *and* (0.7). *If, in addition v_0 has compact support, then*

$$\zeta \in C^1(\mathbb{R}^+).$$

Lacey, Ockendon and Tayler [15] have recently constructed a class of similarity solutions of problem (0.4) which can be used as comparison functions in Corollary C.1. Using their results, one can replace the compact support condition in Corollary C.2 by a polynomial growth condition at $x = -\infty$. Specifically, it suffices to have $v_0 = O(|x|^{2-\beta})$ as $x \downarrow -\infty$ for some $\beta \in \mathbb{R}^+$.

The following example illustrates the scope and limitations of our results. Let

$$v_0(x) = \begin{cases} (1-\theta)\sin^2 x + \theta \sin^4 x & \text{for } x \in [-\pi, 0], \\ 0 & \text{for } x \notin [-\pi, 0], \end{cases}$$

where $\theta \in [0,1]$ is a parameter. It is not difficult to verify that v_0 satisfies the conditions of Corollaries A.1 and C.2 with $\alpha = 1-\theta$, provided that $\theta \in [0, \frac{1}{4}]$. Therefore for $\theta \in [0, \frac{1}{4}]$ we have $t^* = t_m/(1-\theta)$ and $\zeta \in C^1(\mathbb{R}^+)$. However, if $\theta > \frac{1}{4}$, then there is no $\alpha \in \mathbb{R}^+$ such that

$$v_0(x) \le \alpha x^2 \quad \text{in } \mathbb{R}^- \quad \text{and} \quad v_0(x) = \alpha x^2 + o(x^2) \quad \text{as } x \uparrow 0,$$

and we are unable to calculate t^*. Thus, in particular, we cannot apply Theorem C or its corollaries for $\theta > \frac{1}{4}$, and the smoothness of the interface remains an open question. Of course, Theorem A can still be applied to obtain bounds for t^*. For example, if $\theta = 1$, then $\alpha = 0$ and

$$t^* \ge 1.904538 \cdots t_m.$$

We have recently received a preprint from W. L. Kath and D. S. Cohen, *Waiting-time behavior in a nonlinear diffusion equation*, in which they do a formal asymptotic analysis for $m-1$ small which leads to an estimate for the waiting time t^* up to terms which are $o(m-1)$. It would be useful to have more examples in which the waiting time was known accurately. In particular, good numerical schemes for the determination of t^* are sorely needed.

1. Behavior near a vertical interface. Our first result shows that a solution v of (0.3) with a vertical right-hand interface is dominated in $\mathbb{R}^-$ by a multiple of x^2 for each fixed $t \in (0, t^*)$.

LEMMA 1.1. *Let v be a solution of (0.4) in $\mathbb{R} \times (0,T)$ such that $t^* \in (0,T)$. For each $\tau \in (0, t^*)$ there exists a constant $\bar{C} = \bar{C}(\tau)$ such that*

$$v(x,t) \le \frac{\bar{C} x^2}{t^* - t} \quad \text{in } \mathbb{R}^- \times [\tau, t^*). \tag{1.1}$$

Proof. Suppose that (1.1) does not hold. Then for each integer $n \ge 1$ there exists $(x_n, t_n) \in \mathbb{R}^- \times [\tau, t^*)$ such that

$$v(x_n, t_n) > \frac{n x_n^2}{t^* - t_n}. \tag{1.2}$$

Since

$$\int_{x_n}^0 \{-v_x(x,t_n)\}\, dx = v(x_n,t_n) > \frac{n x_n^2}{t^* - t_n},$$

it follows that

$$\frac{1}{-x_n} \int_{x_n}^0 \{-v_x(x,t_n)\}\, dx > \frac{n(-x_n)}{t^* - t_n}.$$

Thus there exists $\xi_n \in (x_n, 0)$ such that

$$-v_x(\xi_n, t_n) > \frac{n(-x_n)}{t^* - t_n}. \tag{1.3}$$

Set $x = -x_n x'$ and $t = (t^* - t_n)t' + t_n$, and define

$$v^*(x',t') = \frac{t^* - t_n}{x_n^2} v(-x_n x', (t^* - t_n)t' + t_n).$$

Observe that v^* is a solution of (0.3) in the variables x', t'. Now

$$v^*_{x'}(x',t')=\frac{t^*-t_n}{-x_n}v_x(-x_nx',(t^*-t_n)t'+t_n)$$

and

$$v^*_{x'x'}(x',t')=(t^*-t_n)v_{xx}(-x_nx',(t^*-t_n)t'+t_n).$$

In view of (1.3),

$$v^*_{x'}\left(-\frac{\xi_n}{x_n},0\right)<-n. \tag{1.4}$$

By the Aronson–Benilan estimate [4],

$$v^*_{x'x'}(x',0)\geq-\frac{t^*-t_n}{t_n}k\geq-\frac{t^*}{\tau}k\equiv-k(\tau). \tag{1.5}$$

For $x'<-\xi_n/x_n$ we have, according to (1.5),

$$-k(\tau)\left(-\frac{\xi_n}{x_n}-x'\right)\leq\int_{x'}^{-\xi_n/x_n}v^*_{x'x'}(z,0)\,dz=v^*_{x'}\left(-\frac{\xi_n}{x_n},0\right)-v^*_{x'}(x',0).$$

Thus, using (1.4),

$$v^*_{x'}(x',0)\leq-n+k(\tau)\left(-\frac{\xi_n}{x_n}-x'\right).$$

Integrating this inequality and using the fact that $v^*(-\xi_n/x_n,0)>0$, we obtain

$$v^*(x',0)\geq n\left(-\frac{\xi_n}{x_n}-x'\right)\left\{1-\frac{k(\tau)}{2}\left(-\frac{\xi_n}{x_n}-x'\right)\right\}$$

for $x'<-\xi_n/x_n$. In particular,

$$v^*(x',0)\geq\frac{n}{2}\left(-\frac{\xi_n}{x_n}-x'\right)\quad\text{for }-\frac{\xi_n}{x_n}-\frac{1}{k(\tau)}\leq x'\leq-\frac{\xi_n}{x_n}. \tag{1.6}$$

Consider the Barenblatt–Pattle solution of (0.3) given by

$$B^*(x',t')=\frac{\beta t_m}{\alpha^2(\beta t'+1)^{(m-1)/(m+1)}}\left\{1-\frac{\alpha^2x'^2}{(\beta t'+1)^{2/m+1}}\right\}^+$$

for constants $\alpha\in\mathbb{R}^+$ and $\beta\in\mathbb{R}^+$. If we choose α and β so that

$$\operatorname{supp}B^*(x',0)=\left[-\frac{1}{2k(\tau)},\frac{1}{2k(\tau)}\right],\qquad B^*_{x'}\left(\frac{1}{2k(\tau)},0\right)=-\frac{n}{2}$$

then

$$\alpha=2k(\tau)\quad\text{and}\quad\beta=\frac{nk(\tau)}{2t_m}.$$

Observe that the right-hand interface for $B^*(x',t')$ moves one unit to the right in $4t_m/n$ units of time.

In view of (1.6),

$$v^*(x',0)\geq B^*\left(x'+\frac{\xi_n}{x_n}+\frac{1}{2k(\tau)},0\right)\quad\text{in }\mathbb{R}.$$

Hence by the comparison principle[1]

$$v^*(x',t') \geq B^*\left(x' + \frac{\xi_n}{x_n} + \frac{1}{2k(\tau)}, t'\right) \quad \text{in } \mathbb{R}\times(0,1). \tag{1.7}$$

The right-hand interface for v^* is vertical for $t\in(0,1)$. On the other hand, if $n>4t_m$ then the right-hand interface for B^* is to the right of $-\xi_n/x_n+1>0$ for $t=1$. Thus the right-hand interfaces for v^* and B^* cross in the interval (0.1), and this contradicts (1.7). □

The next result contains Theorem A, as well as additional information about the behavior of solutions near a vertical interface which will be used in the proof of Theorem C.

PROPOSITION 1.2. *Let v be a solution of problem* (0.4) *where v_0 satisfies*

$$v_0(x)=\alpha x^2+o(x^2) \quad \text{as } x\uparrow 0, \quad v_0(x)\leq \beta x^2 \quad \text{in } \mathbb{R}^-, \quad \text{and} \quad v_0(x)=0 \quad \text{in } \mathbb{R}^+$$

for some constants $\alpha\in\mathbb{R}^+$ and $\beta\in\mathbb{R}^+$. Then

$$\frac{t_m}{\beta}\leq t^*\leq\frac{t_m}{\alpha}$$

and

$$v(x,t)=\frac{t_m x^2}{(t_m/\alpha)-t}+o(x^2) \quad \text{as } x\uparrow 0$$

uniformly for t in any compact subset of $(0,t^*)$.

Proof. Since $v_0\leq\beta x^2$ in $\mathbb{R}^-$ and $v_0=0$ in $\mathbb{R}^+$, it follows that

$$v(x,t)\begin{cases}\leq\dfrac{t_m x^2}{(t_m/\beta)-t} & \text{in } (-\infty,0]\times[0,t_m/\beta),\\ =0 & \text{in } \mathbb{R}^+\times[0,t_m/\beta).\end{cases} \tag{1.8}$$

Thus, in particular, $t^*\geq t_m/\beta$. According to Lemma 1.1,

$$v(x,t)\begin{cases}\leq\dfrac{\bar{C}(\tau)x^2}{t^*-t} & \text{in } (-\infty,0]\times[\tau,t^*),\\ =0 & \text{in } \mathbb{R}^+\times[\tau,t^*)\end{cases} \tag{1.9}$$

for any $\tau\in(0,t^*)$. Set $\tau=t_m/2\beta$. Then, in view of (1.8) and (1.9), for each $\varepsilon\in(0,t^*)$ there exists a constant $C_1(\varepsilon)\in\mathbb{R}^+$ such that

$$v(x,t)\begin{cases}\leq C_1(\varepsilon)x^2 & \text{in } (-\infty,0]\times[0,t^*-\varepsilon],\\ =0 & \text{in } \mathbb{R}^+\times[0,t^*-\varepsilon].\end{cases} \tag{1.10}$$

For $\delta\in\mathbb{R}^+$ define

$$v_\delta(x,t)\equiv\frac{1}{\delta^2}v(\delta x,t).$$

---

[1] Note that v^* is simply a weak solution of (0.3); in particular, there are no a priori global assumptions on its initial values or on its behavior as $x'\to-\infty$. Thus the application of the comparison principle in the present circumstances must be justified. This is done in detail in [5].

Note that v_δ is a solution of (0.3) and that, for each $\varepsilon\in(0,t^*)$, v_δ satisfies (1.9) and (1.10) for all $\delta\in\mathbb{R}^+$. Moreover, $v_0(x)=v(x,0)=\alpha x^2+o(x^2)$ in $(-l,0)$ for some $l\in\mathbb{R}^+$ implies that

$$v_\delta(x,0)=\alpha x^2\{1+o(1)\} \quad \text{in } \left(-\frac{l}{\delta},0\right) \tag{1.11}$$

as $\delta\to 0$.

Fix $\varepsilon\in(0,t^*)$. For each integer $n\geq 1$ such that $1/n<t^*-\varepsilon$, let

$$S_n=[-n,n]\times\left[\frac{1}{n},t^*-\varepsilon\right]$$

and

$$S_n^*=\left[-n-\frac{1}{2},n+\frac{1}{2}\right]\times\left[\frac{1}{2n},t^*-\varepsilon\right].$$

Let $\{\delta_k\}$ be a sequence such that $\delta_k\to 0$ as $k\to\infty$. According to (1.10), for each fixed n, the sequence $\{v_{\delta_k}\}$ is uniformly bounded in S_n^*. Therefore, as shown in [1], the sequence $\{v_{\delta_k}\}$ is Lipschitz continuous with respect to x uniformly in S_n. Moreover, by the results of [14], the sequence $\{v_{\delta_k}\}$ is Hölder continuous as a function of (x,t) uniformly in S_n.

Define $u_k=\{(m-1)v_{\delta_k}/m\}^{1/(m-1)}$. As we have shown above, the sequence $\{u_k\}$ is uniformly bounded and equicontinuous in S_n. Note that

$$\frac{\partial}{\partial x}u_k^m=u_k\frac{\partial}{\partial x}v_{\delta_k}.$$

In view of the uniform Lipschitz continuity of v_{δ_k}, there exists a constant $c_n>0$ such that

$$\left|\frac{\partial}{\partial x}v_{\delta_k}\right|\leq c_n \quad \text{in } S_n.$$

Therefore, the sequence $\{u_k^m\}$ is weakly compact in $L^2[1/n,t^*-\varepsilon;H^{1,2}(-n,n)]$. Thus, from each sequence $\{k_i\}$ we can extract a subsequence $\{k_j^n\}$ such that $k_j^n\to\infty$,

$$u_{k_j^n}\to\tilde{u} \quad \text{uniformly in } S_n,$$

and

$$\frac{\partial}{\partial x}u_{k_j^n}^m\to\frac{\partial}{\partial x}\tilde{u}^m \quad \text{weakly in } L^2\left[1/n,t^*-\varepsilon;H^{1,2}(-n,n)\right]$$

as $j\to\infty$. By the usual diagonal procedure we can construct a sequence $\{l_i\}$ such that $l_i\to\infty$,

$$u_{l_i}\to\tilde{u} \quad \text{in } \mathbb{R}\times(0,t^*-\varepsilon], \tag{1.12}$$

and

$$\frac{\partial}{\partial x}u_{l_i}^m\to\frac{\partial}{\partial x}\tilde{u}^m \quad \text{weakly in } L^2_{\text{loc}}\left(0,t^*-\varepsilon;H^{1,2}_{\text{loc}}(\mathbb{R})\right) \tag{1.13}$$

as $i\to\infty$. Moreover, the convergence in (1.12) is uniform on compact subsets of $\mathbb{R}\times(0,t^*-\varepsilon]$.

Each u_k satisfies

$$\int_0^{t^*-\varepsilon}\int_{\mathbb{R}}\left\{\frac{\partial\varphi}{\partial x}\frac{\partial u_k^m}{\partial x}-\frac{\partial\varphi}{\partial t}u_k\right\}dx\,dt-\int_{\mathbb{R}}u_k(x,0)\varphi(x,0)\,dx=0$$

for all $\varphi \in C^\infty(\mathbb{R}\times[0,T])$ which vanish for $|x|$ large and for t near $t^*-\varepsilon$. Fix $\eta \in (0,t^*-\varepsilon)$. Then, integrating by parts, we obtain

$$\int_\eta^{t^*-\varepsilon}\int_{\mathbb{R}}\left\{\frac{\partial\varphi}{\partial x}\frac{\partial u_k^m}{\partial x}-\frac{\partial\varphi}{\partial t}u_k\right\}dx\,dt-\int_{\mathbb{R}}u_k(x,0)\varphi(x,0)\,dx=\int_0^\eta\int_{\mathbb{R}}\left\{\frac{\partial^2\varphi}{\partial x^2}u_k^m+\frac{\partial\varphi}{\partial t}u_k\right\}dx\,dt.$$

Since $\operatorname{supp}\varphi(\cdot,t)$ is bounded for $t\in[0,T]$ and, in view of (1.10), u_k is bounded on compact subsets of $\mathbb{R}\times[0,t^*-\varepsilon]$, there exists a constant $\Phi>0$ depending only on φ such that

$$\left|\int_\eta^{t^*-\varepsilon}\int_{\mathbb{R}}\left\{\frac{\partial\varphi}{\partial x}\frac{\partial u_k^m}{\partial x}-\frac{\partial\varphi}{\partial t}u_k\right\}dx\,dt-\int_{\mathbb{R}}u_k(x,0)\varphi(x,0)\,dx\right|\le\eta\Phi.$$

If we set $k=l_i$ and let $i\to\infty$, then it follows from (1.11), (1.12), and (1.13) that

$$\left|\int_\eta^{t^*-\varepsilon}\int_{\mathbb{R}}\left\{\frac{\partial\varphi}{\partial x}\frac{\partial\tilde u^m}{\partial x}-\frac{\partial\varphi}{\partial t}\tilde u\right\}dx\,dt-\int_{\mathbb{R}}\left(\frac{m-1}{m}\alpha x^2\right)^{1/(m-1)}\varphi(x,0)\,dx\right|\le\eta\Phi.$$

Now let $\eta\to 0$ to obtain

$$\int_0^{t^*-\varepsilon}\int_{\mathbb{R}}\left\{\frac{\partial\varphi}{\partial x}\frac{\partial\tilde u^m}{\partial x}-\frac{\partial\varphi}{\partial t}\tilde u\right\}dx\,dt-\int_{\mathbb{R}}\left(\frac{m-1}{m}\alpha x^2\right)^{1/(m-1)}\varphi(x,0)\,dx=0.$$

By Kalashnikov's uniqueness theorem [11],

$$\tilde u(x,t)=\left\{\frac{(m-1)t_m x^2}{m((t_m/\alpha)-t)}\right\}^{1/(m-1)}.$$

Therefore, we conclude that

$$\lim_{\delta\downarrow 0}v_\delta(x,t)=\frac{t_m x^2}{(t_m/\alpha)-t}\quad\text{on }(-\infty,0]\times(0,t^*-\varepsilon], \tag{1.14}$$

where the convergence is uniform on compact subsets.

Suppose that $t_m/\alpha<t^*$. Set $\tau=t_m/2\alpha$. Then, according to (1.9),

$$v_\delta(x,t)\le\frac{\bar C(\tau)x^2}{t^*-t}\quad\text{in }(-\infty,0]\times[\tau,t^*).$$

Thus, if $\varepsilon\in(t^*-(t_m/\alpha),t^*-\tau)$ it follows from (1.14) that

$$\frac{t_m}{(t_m/\alpha)-t^*+\varepsilon}\le\frac{\bar C(\tau)}{\varepsilon}.$$

If we let $\varepsilon\downarrow t^*-(t_m/\alpha)$, this leads to a contradiction, so we conclude that $t_m/\alpha\ge t^*$.

Fix $\varepsilon\in(0,t^*)$ and $t_0\in(0,t^*-\varepsilon)$. For each $\eta>0$ there is a $\delta_0=\delta_0(\varepsilon,\eta,t_0)>0$ such that $\delta\in(0,\delta_0)$ implies

$$\left|v_\delta(-1,t)-\frac{t_m}{(t_m/\alpha)-t}\right|<\eta\quad\text{for }t\in[t_0,t^*-\varepsilon].$$

In view of the definition of v_δ it follows that

$$\left|v(-\delta,t)-\frac{t_m\delta^2}{(t_m/\alpha)-t}\right|<\eta\delta^2\quad\text{for }t\in[t_0,t^*-\varepsilon],\quad\delta\in(0,\delta_0).$$

Therefore

$$v(x,t)=\frac{t_m x^2}{(t_m/\alpha)-t}+o(x^2) \quad \text{as } x\uparrow 0$$

uniformly on compact subsets of $(0,t^*-\varepsilon]$. □

2. Behavior near a nonvertical interface. In this section we show that in the neighborhood of a point where the interface is moving we have $v\sim L_\gamma$ where L_γ is the linear solution of equation (0.3) described in the Introduction. To accomplish this we need the following technical lemmas.

LEMMA 2.1. *Let v be a solution of* (0.3). *If for some $t_0\in(0,T)$, $x_0=\zeta(t_0)$ and*

$$\lim_{x\uparrow x_0} -v_x(x,t_0)=\gamma>0, \tag{2.1}$$

then there exists constants C_1, C_2, δ_0, and $A\in\mathbb{R}^+$ depending on γ, t_0, and T such that

$$C_1(x_0-x)\le v(x,t)\le C_2(x_0-x)$$

for all (x,t) such that

$$0\le t_0-t<A(x_0-x)<\delta_0.$$

Proof. By Taylor's theorem, for $t\in(0,t_0)$ and $x<\zeta(t)$,

$$v(x,t)=\{\zeta(t)-x\}\{-v_x(\tilde{x},t)\}$$

where $x<\tilde{x}<\zeta(t)$. If $t\ge\eta>0$ then there exists a constant $C_2=C_2(\eta)\in\mathbb{R}^+$ such that [1]

$$-v_x(\tilde{x},t)\le C_2(\eta).$$

Then since $\zeta(t)\le x_0$, it follows that

$$v(x,t)\le C_2(\eta)(x_0-x) \quad \text{for } (x,t)\in(-\infty,x_0]\times[\eta,t_0]. \tag{2.2}$$

In view of (2.1),

$$\lim_{h\downarrow 0}\frac{\zeta(t_0+h)-\zeta(t_0)}{h}=\gamma.$$

Thus, there exists a $\nu\in\mathbb{R}^+$ such that

$$x_0+\frac{\gamma}{2}(t-t_0)\le\zeta(t)\le x_0+\frac{3\gamma}{2}(t-t_0) \quad \text{for } t\in[t_0,t_0+\nu]. \tag{2.3}$$

For $\varepsilon\in(0,\nu]$, define

$$v_\varepsilon(x,t)=\begin{cases}\dfrac{t_m\left\{x-\left(x_0+\frac{\varepsilon}{2}\gamma\right)\right\}^2}{t_0+\varepsilon-t} & \text{in } \left(-\infty,x_0+\frac{\varepsilon}{2}\gamma\right]\times[t_0-\varepsilon,t_0+\varepsilon),\\ 0 & \text{in } \left(x_0+\frac{\varepsilon}{2}\gamma,+\infty\right)\times[t_0-\varepsilon,t_0+\varepsilon).\end{cases}$$

Then v_ε is a solution of (0.3) with interface

$$x=\zeta_\varepsilon(t)\equiv x_0+\frac{\varepsilon}{2}\gamma.$$

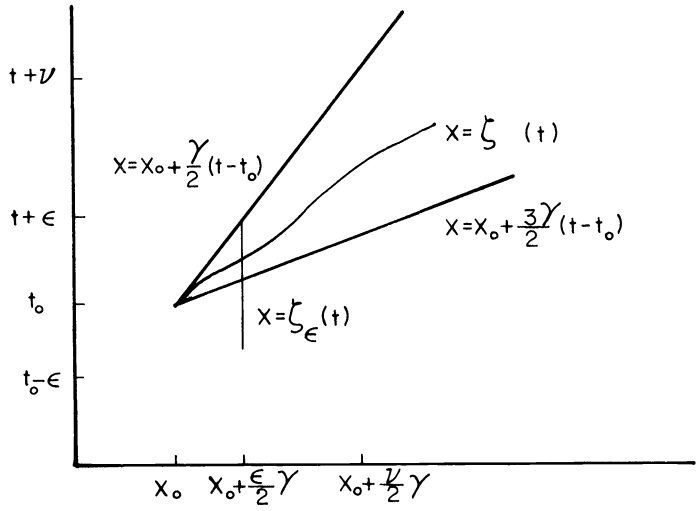


FIG. 1

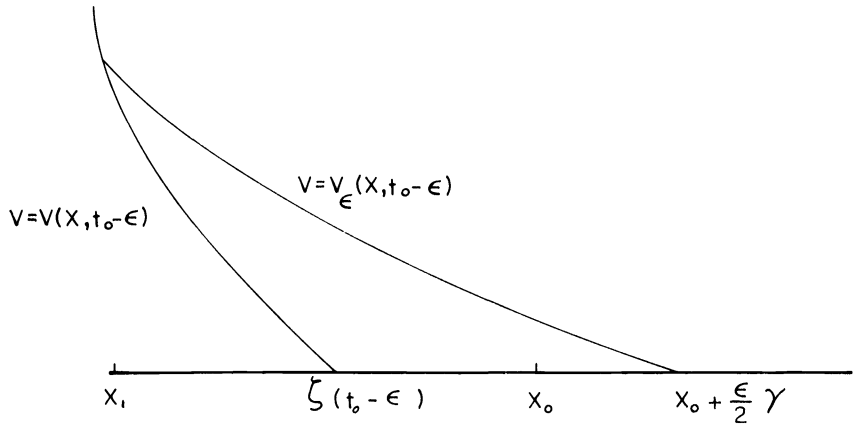


FIG. 2

In view of (2.3), the interfaces $x=\zeta(t)$ and $x=\zeta_\varepsilon(t)$ intersect somewhere on the interval $(t_0+\varepsilon/3, t_0+\varepsilon)$ (cf. Fig. 1). Thus we cannot have

$$v(x,t_0-\varepsilon)\le v_\varepsilon(x,t_0-\varepsilon) \quad \text{for all } x\le\zeta(t_0-\varepsilon).$$

That is, there exists $x_1=x_1(\varepsilon)<\zeta(t_0-\varepsilon)$ such that (cf. Fig. 2)

$$v(x,t_0-\varepsilon)<v_\varepsilon(x,t_0-\varepsilon) \quad \text{for } x\in(x_1,\zeta(t_0-\varepsilon)],$$

and

$$v(x_1,t_0-\varepsilon)=v_\varepsilon(x_1,t_0-\varepsilon). \tag{2.4}$$

Note that this implies that

$$v_x(x_1,t_0-\varepsilon)\le v_{\varepsilon x}(x_1,t_0-\varepsilon). \tag{2.5}$$

If $t_0-\varepsilon\ge\eta>0$ then in view of (2.4)

$$-C_2(\eta)\le v_x(x_1,t_0-\varepsilon)\le v_{\varepsilon x}(x_1,t_0-\varepsilon)=\frac{t_m(x_1-x_0-\varepsilon\gamma/2)}{\varepsilon}\le 0.$$

Therefore, there exists a constant $C'=C'(\eta)$ such that

$$0\le x_0-x_1\le C'(\eta)\varepsilon.$$

On the other hand, by (2.2),

$$v(x_1,t_0-\varepsilon)\le C_2(\eta)(x_0-x_1),$$

and (2.4) implies

$$C_2(\eta)(x_0-x_1)\ge\frac{t_m}{2\varepsilon}\left(x_0-x_1+\frac{\varepsilon}{2}\gamma\right)^2\ge\frac{\gamma^2 t_m}{\delta}\varepsilon.$$

Thus we have shown that there exist constants $C'(\eta)$, $C''=C''(\eta,\gamma)\in\mathbb{R}^+$ such that

$$0<C''\varepsilon\le x_0-x_1(\varepsilon)<C'\varepsilon \quad \text{for } \varepsilon\in[0,t_0-\eta]. \tag{2.6}$$

By the Aronson–Benilan estimate [4],

$$v_{xx}(\xi,t_0-\varepsilon)\ge-\frac{k}{t_0-\varepsilon}\ge-\frac{k}{\eta}\equiv-\bar{k}(\eta)$$

for $\xi\in\mathbb{R}$ and $\varepsilon\in[0,t_0-\eta]$. Therefore, for $x<x_1$,

$$-\bar{k}(\eta)(x_1-x)\le\int_x^{x_1}v_{xx}(\xi,t_0-\varepsilon)\,d\xi=v_x(x_1,t_0-\varepsilon)-v_x(x,t_0-\varepsilon).$$

In view of (2.5) and (2.6),

$$v_x(x,t_0-\varepsilon)\le-\frac{t_m}{\varepsilon}\left(x_0-x_1+\frac{\varepsilon}{2}\gamma\right)+\bar{k}(\eta)(x_1-x)\le-t_m\left(C''+\frac{\gamma}{2}\right)+\bar{k}(x_1-x).$$

Thus

$$v_x(x,t_0-\varepsilon)\le-\frac{t_m}{2}\left(C''+\frac{\gamma}{2}\right)\equiv-C''', \tag{2.7}$$

provided that

$$x\ge x_0-\frac{t_m}{2\bar{k}}\left(C''+\frac{\gamma}{2}\right)\equiv x_0-B.$$

For $x\in[x_0-B,x_1)$ it follows from (2.7) that

$$-C'''(x_1-x)\ge\int_x^{x_1}v_x(\xi,t_0-\varepsilon)\,d\xi=v(x_1,t_0-\varepsilon)-v(x,t_0-\varepsilon).$$

Thus

$$v(x,t_0-\varepsilon)\ge v(x_1,t_0-\varepsilon)+C'''(x_1-x)\ge C'''(x_1-x).$$

Write

$$x_1-x=(x_0-x)\left\{1-\frac{x_0-x_1}{x_0-x}\right\}.$$

If $\varepsilon<A(x_0-x)<\delta_0$ for some $A\in\mathbb{R}^+$ then

$$x_0-\frac{\delta_0}{A}<x<x_0-\frac{\varepsilon}{A}$$

and

$$1-\frac{x_0-x_1}{x_0-x}>1-AC'.$$

Thus if we take $A=2/C'$ and $\delta_0=2B/C'$ then

$$v(x,t_0-\varepsilon)\geq\frac{C'''}{2}(x_0-x).$$

This together with (2.2) proves the lemma. □

LEMMA 2.2. *Under the hypothesis of Lemma* 2.1 *there exists a constant* $C_3\in\mathbb{R}^+$ *depending only on* γ, t_0, *and* T *such that*

$$|v_{tt}(x,t)|\leq\frac{C_3}{x_0-x}$$

for all (x,t) *which satisfy*

$$0\leq t_0-t<\frac{3}{4}A(x_0-x)\leq\frac{4}{5}\delta_0,$$

where A *and* δ_0 *are as in Lemma* 2.1.

Proof. For $\delta\in(0,4\delta_0/5)$ define

$$v_\delta(x,t)\equiv\frac{1}{\delta}v(\delta x+x_0,\delta t+t_0).$$

Let $D\equiv\{(x,t): Ax<t\leq 0,\ -\frac{5}{4}<x<-\frac{1}{4}\}$. By Lemma 2.1, for $(x,t)\in D$

$$0<\frac{1}{32}C_1<v_\delta(x,t)=\frac{1}{\delta}v(\delta x+x_0,\delta t+t_0)<\frac{5}{4}C_2. \tag{2.8}$$

Thus $w=v_\delta$ is a classical solution of the equation

$$w_t=\{(m-1)v_\delta w_x\}_x+(2-m)v_{\delta x}w_x \tag{2.9}$$

in D. In view of (2.8), equation (2.9) is uniformly parabolic in D. Moreover,

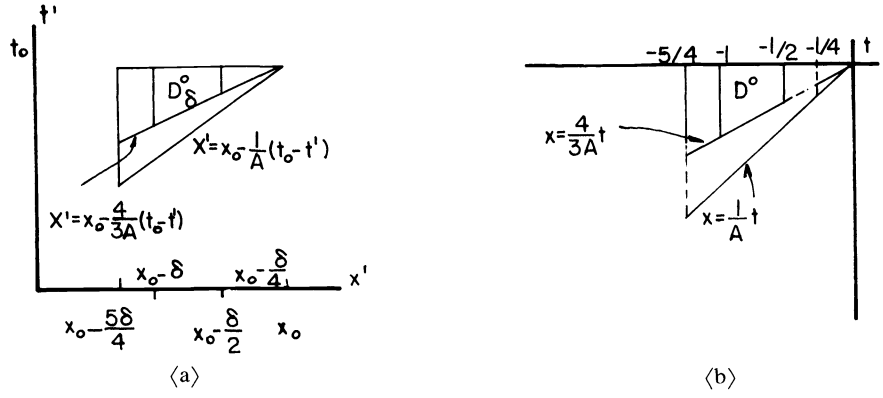
$$|v_{\delta x}(x,t)|=|v_x(\delta x+x_0,\delta t+t_0)|\leq C_2(\eta)$$

for $\delta t+t_0\geq\eta$. Note that these bounds are independent of δ. By the results of [6], v_δ and $v_{\delta x}$ have Hölder norms independent of δ in any compact subset of D. We can now apply the Schauder-type theory for parabolic equations [10] to the equation

$$w_t=(m-1)v_\delta w_{xx}+v_{\delta x}w_x$$

and its derivative with respect to t to conclude that $v_\delta, v_{\delta t}, v_{\delta x}, v_{\delta xx}, v_{\delta tx}, v_{\delta txx}$, and $v_{\delta tt}$ all have Hölder norms independent of δ in (cf. Fig. 3b)

$$D^0=\left\{(x,t): \frac{3A}{4}x\leq t\leq 0,\ -1\leq x\leq-\frac{1}{2}\right\}.$$



⟨a⟩ ⟨b⟩

FIG. 3

Thus, in particular there exists a constant C_3 depending only on γ, t_0 and T such that

$$|v_{\delta tt}(x,t)| = \delta |v_{tt}(\delta x + x_0, \delta t + t_0)| \le C_3 \quad \text{in } D^0.$$

If $(x,t) \in D^0$, then

$$(x',t') \in D_\delta^0 \equiv \left\{ (x',t') : t_0 - \frac{3A}{4}(x_0 - x') \le t' \le t_0,\ x_0 - \delta \le x' \le x_0 - \frac{\delta}{2} \right\}$$

where $x' = \delta x + x_0$ and $t' = \delta t + t_0$ (cf. Fig. 3a). Since

$$\delta = \frac{x' - x_0}{x} = \frac{x_0 - x'}{-x} \ge x_0 - x',$$

it follows that

$$|v_{tt}(x',t')| \le \frac{C_3}{x_0 - x'} \quad \text{in } D_\delta^0.$$

Finally, since $\delta \in (0, 4\delta_0/5)$ is arbitrary and C_3 is independent of δ, this proves the assertion. □

PROPOSITION 2.3. *Let v be a solution of* (0.3). *If for some $t_0 \in (0,T)$*

$$\lim_{x \uparrow x_0} -v_x(x,t_0) = \gamma > 0,$$

where $x_0 = \zeta(t_0)$, then in a neighborhood of (x_0,t_0)

$$v(x,t) = L_\gamma(x - x_0, t - t_0) + o\big(|x - x_0| + |t - t_0|\big)$$

where $L_\gamma(x,t) = (\gamma^2 t - \gamma x)^+$.

Proof. Let v_δ be as defined in Lemma 2.2. Fix $\eta \in (0, t_0 \wedge T - t_0)$. Then v is defined in the η-neighborhood of (x_0,t_0) given by

$$(x_0 - \eta, x_0 + \eta) \times (t_0 - \eta, t_0 + \eta),$$

and v_δ is defined in the corresponding η/δ-neighborhood of the origin.

Observe that v_δ is a solution of equation (0.3) satisfying

$$|v_{\delta x}(x,t)| \le C_2(t_0 - \eta)$$

and [4]

$$v_{\delta xx}(x,t) = \delta v_{xx}(\delta x + x_0, \delta t + t_0) \ge -\frac{\delta k}{\delta t + t_0} \ge -\frac{\delta k}{t_0 - \eta} \tag{2.10}$$

for $t \ge -\eta/\delta$. According to the results of [1] and [14], v_δ is Lipschitz continuous with respect to x and Hölder continuous with respect to t. Therefore, for each sequence $\{\delta_n\}$ such that $\delta_n \downarrow 0$, there exists a subsequence $\{\delta_{n'}\}$ such that $\delta_{n'} \downarrow 0$ and $v_{\delta_{n'}} \to v^*$ in $\mathbb{R}^2$ uniformly on compact sets. Moreover, v^* is a weak solution of the pressure equation and all of the derivatives of $v_{\delta_{n'}}$ converge to the corresponding derivatives of v^* on $\{(x,t) : v^*(x,t) > 0\}$.

In view of our hypothesis, for each fixed $x \in \mathbb{R}^-$

$$v_x^*(x,0) = \lim_{n' \to \infty} v_{\delta_{n'} x}(x,0) = \lim_{n' \to \infty} v_x(\delta_{n'} x + x_0, t_0) = -\gamma,$$

and for each fixed $x \ge 0$

$$v^*(x,0) = \lim_{n' \to \infty} v_{\delta_{n'}}(x,0) = \lim_{n' \to \infty} \frac{1}{\delta_{n'}} v(\delta_{n'} x + x_0, t_0) = 0.$$

Thus

$$v^*(x,0)=\begin{cases} -\gamma x & \text{for } x\le 0,\\ 0 & \text{for } x\ge 0.\end{cases}$$

Moreover,

$$v_\delta(x,t)\le\frac{C_2}{\delta}|\delta x+x_0-\zeta(\delta t+t_0)|\le\frac{C_2}{\delta}\left(\delta|x|+|x_0-\zeta(\delta t+t_0)|\right)$$
$$\le C_2\left(|x|+\frac{3\gamma}{2}|t|\right)$$

implies

$$v^*(x,t)\le C_2\left(|x|+\frac{3\gamma}{2}|t|\right).$$

In particular, for any $T\in\mathbb{R}^+$, v^* grows no faster than linearly in $\mathbb{R}\times(0,T)$. Therefore, in view of Kalashnikov's uniqueness theorem [11],

$$v^*(x,t)=L_\gamma(x,t)\quad\text{in } \mathbb{R}\times[0,+\infty).$$

By Lemma 2.2,

$$|v_{\delta tt}(x,t)|=\delta|v_{tt}(\delta x+x_0,\delta t+t_0)|\le\frac{C_3}{-x}$$

for $0\le -t<-3Ax/4<4\delta_0/5\delta$. Therefore

$$|v^*_{tt}(x,t)|\le\frac{C_3}{-x}\quad\text{for } \frac{3A}{4}x<t\le 0.$$

For $t<0$ and $x<\gamma t$,

$$v^*(x,t)=v^*(x,0)+tv^*_t(x,0)+\frac{t^2}{2}v^*_{tt}(x,\tilde t)=L_\gamma(x,0)+tL_{\gamma t}(x,0)+\frac{t^2}{2}v^*_{tt}(x,\tilde t),$$

where $t<\tilde t<0$. Thus

$$v^*(x,t)-L_\gamma(x,t)=\frac{t^2}{2}v^*_{tt}(x,\tilde t),$$

and it follows that

$$\lim_{x\downarrow-\infty}\left\{v^*(x,t)-L_\gamma(x,t)\right\}=0\quad\text{for } t\in\mathbb{R}^-. \tag{2.11}$$

Next we show that

$$v^*(x,t)\ge L_\gamma(x,t)\quad\text{for } t\in\mathbb{R}^-.$$

To prove this, first observe that (2.11) implies that $v^*_x(x,t)\ge-\gamma$. Suppose for contradiction that for some $\tilde x<\zeta(t)$ and $\varepsilon>0$ we have

$$v^*_x(\tilde x,t)=-\gamma-\varepsilon.$$

In view of (2.10), $v^*_{xx}\ge 0$. Thus for $x<\tilde x$ and some $\hat x\in(x,\tilde x)$

$$v^*(x,t)=v^*(\tilde x,t)+(x-\tilde x)v^*_x(\hat x,t)\ge v^*(\tilde x,t)+(\tilde x-x)(\gamma+\varepsilon).$$

It follows that

$$v^*(x,t)-L_\gamma(x,t)\ge v^*(\tilde x,t)+\tilde x(\gamma+\varepsilon)-\gamma^2t-\varepsilon x$$

and, letting $x \downarrow -\infty$, this contradicts (2.11). The assertion now follows from (2.11) since $L_{\gamma x} = -\gamma$ so that for each fixed $x_0 \in \mathbb{R}$

$$v^*(x_0,t) - L_\gamma(x_0,t) \geq v^*(x,t) - L_\gamma(x,t)$$

for all $x < x_0$.

Finally we show that

$$v^*(x,t) = L_\gamma(x,t) \quad \text{for } t \in \mathbb{R}^-.$$

Consider a fixed $(\tilde{x},\tilde{t}) \in \mathbb{R}^-$ such that $\tilde{x} \leq \gamma\tilde{t}$. Let N be a closed neighborhood of $(\tilde{x},\tilde{t})$ such that $(x,t) \in N$ implies that $x \leq \gamma t$ and N contains points with $t > 0$. In N, $v^* - L_\gamma \geq 0$ and achieves its minimum value, 0, on $N \cap \{t > 0\}$. By the strong minimum principle $v^* \equiv L_\gamma$ in N. Therefore

$$v^*(x,t) = L_\gamma(x,t) \quad \text{in } \{(x,t) \in \mathbb{R} \times \mathbb{R}^- : x \leq \gamma t\}.$$

Suppose that $v^*(\tilde{x},\tilde{t}) > 0$ for some $(\tilde{x},\tilde{t}) \in \mathbb{R} \times \mathbb{R}^-$ with $x > \gamma t$. By [12], $v(\tilde{x},t) > 0$ for all $t > \tilde{t}$. However, this contradicts the fact that the line $x = \tilde{x}$ must intersect the line $x = \gamma t$ for some $t > \tilde{t}$ and $v = L_\gamma = 0$ at that intersection.

Since $v_\delta \to v^* \equiv L_\gamma$, given $\varepsilon > 0$ there exists a $\delta_0 = \delta_0(\varepsilon) > 0$ such that $\delta < \delta_0$ implies that

$$\left| \frac{1}{\delta} v(\delta + x_0, \delta + t_0) - L_\gamma(1,1) \right| < \varepsilon.$$

Set $x = \delta + x_0$ and $t = \delta + t_0$. Then

$$\delta L_\gamma(1,1) = L_\gamma(x - x_0, t - t_0)$$

and

$$|v(x,t) - L_\gamma(x - x_0, t - t_0)| < \varepsilon\delta.$$

The assertion now follows, since $|\delta| = \frac{1}{2}\{|x - x_0| + |t - t_0|\}$. □

3. Smoothness. In this section we shall prove Theorem B as well as Theorem C and its corollaries. Theorem B was first proved by Caffarelli and Friedman in [8].

If v is a solution of problem (0.4) and v_0 satisfies the hypothesis of Theorem B, then the right-hand interface $\zeta(t)$ is a nondecreasing Lipschitz continuous function on $(0,T)$ (see [2]). Thus ζ' exists almost everywhere on $(0,T)$ and, as was shown in [13],

$$D^+\zeta(t) = -v_x(\zeta(t),t) \tag{3.1}$$

everywhere in $(0,T)$. Caffarelli and Friedman [8] prove that for each $\delta \in (0,T)$, there is constant $K > 0$, depending only on m and the lower bound for v_{xx} in $\mathbb{R} \times [\delta,T)$, and a positive measure μ such that

$$\zeta'' + K\zeta' = \mu \tag{3.2}$$

in the sense of distributions on (δ,T). From (3.2) they derive the representation formula

$$\zeta = \eta + \xi, \tag{3.3}$$

where η is Lipschitz continuously differentiable and ζ is convex on (δ,T). In view of this representation, to prove that $\zeta \in C^1(\delta,T)$ for some $\delta > 0$ it suffices to show that ζ and hence ξ is differentiable everywhere on (δ,T). This follows, since a convex function which is differentiable everywhere on an interval is necessarily continuously differentiable there. Thus to prove Theorem B it suffices to show that $D^+\zeta(\tau) > 0$ for some

$\tau\in(0,T)$ implies that ζ is differentiable on (τ,T). Here we shall give a new and somewhat simplified proof of this fact based on (3.2), (3.3), and Proposition 2.3.

Since $D^+\zeta=\zeta'$ almost everywhere in (δ,T), (3.2) implies that

$$(e^{Kt}D^+\zeta)'\geq 0 \tag{3.4}$$

in the sense of distributions on (δ,T). In view of the representation formula (3.3), $D^+\zeta(t+0)$ and $D^+\zeta(t-0)$ exist for all $t\in(\delta,T)$ and satisfy

$$D^+\zeta(t+0)\geq D^+\zeta(t)\geq D^+\zeta(t-0).$$

By a standard argument, it follows from (3.4) that

$$e^{Kt_2}D^+\zeta(t_2-0)\geq e^{Kt_1}D^+\zeta(t_1+0)$$

for $\delta<t_1<t_2<T$. Thus, in particular, $\delta<t_1<t_2<T$ implies

$$D^+\zeta(t_2)\geq e^{K(t_1-t_2)}D\zeta^+(t_1). \tag{3.5}$$

Proof of Theorem B. We assume that $D^+\zeta(\tau)>0$ for some $\tau\in(0,T)$ and fix $\delta\in(0,\tau)$. For arbitrary $t_0\in(\tau,T)$, set $x_0=\zeta(t_0)$ and $\gamma=D^+\zeta(t_0)$. In view of (3.5),

$$D^+\zeta(t_0)\geq e^{K(\tau-t_0)}D^+\zeta(\tau)>0. \tag{3.6}$$

We shall show that $D^-\zeta(t_0)=D^+\zeta(t_0)$.

If $\gamma'>\gamma$, then there exists $\varepsilon'\in(0,t_0-\tau)$ such that

$$\zeta(\tau)>x_0+\gamma'(t-t_0) \tag{3.7}$$

for all $t\in(t_0-\varepsilon',t_0)$. Otherwise there exists a sequence $\{\varepsilon_n\}$ such that $\varepsilon_n\downarrow 0$ and

$$\zeta(t_0-\varepsilon_n)\leq x_0-\varepsilon_n\gamma'.$$

By the definition of the interface $\zeta(t)$ and Proposition 2.3,

$$0=v(x_0-\varepsilon\gamma',t_0-\varepsilon_n)=L_\gamma(-\varepsilon_n\gamma',-\varepsilon_n)+o(\varepsilon_n)=-\gamma^2\varepsilon_n+\gamma\gamma'\varepsilon_n+o(\varepsilon_n).$$

Thus

$$0=\gamma(\gamma'-\gamma)+o(1)\quad\text{as } n\to\infty,$$

which contradicts $\gamma'-\gamma>0$. Therefore (3.7) holds and we conclude that

$$\limsup_{t\uparrow t_0}\frac{x_0-\zeta(t)}{t_0-t}\leq\gamma'.$$

Since $\gamma'>\gamma$ is arbitrary, it follows that

$$\limsup_{t\to t_0}\frac{\zeta(t_0)-\zeta(t)}{t_0-t}\leq\gamma. \tag{3.8}$$

If $\gamma''<\gamma$ then there exists an $\varepsilon''\in(0,t_0-\tau)$ such that

$$\zeta(x)<x_0+\gamma''(t-t_0) \tag{3.9}$$

for all $t\in(t_0-\varepsilon'',t_0)$. For otherwise there exists a sequence $\{\varepsilon_n\}$ such that $\varepsilon_n\downarrow 0$ and

$$\zeta(t_0-\varepsilon_n)\geq x_0-\gamma''\varepsilon_n>x_0-\gamma\varepsilon_n.$$

By the theorem of the mean there exists $\tilde{x}\in(x_0-\varepsilon_n\gamma,x_0-\varepsilon_n\gamma'')$ such that

$$v_x(\tilde{x},t_0-\varepsilon_n)=\{v(x_0-\gamma''\varepsilon_n,t_0-\varepsilon_n)-v(x_0-\gamma\varepsilon_n,t_0-\varepsilon_n)\}/\varepsilon_n(\gamma-\gamma'').$$

Since $\gamma'' < \gamma$ it follows from Proposition 2.3 that

$$v_x(\tilde{x}, t_0 - \varepsilon_n) = o(\varepsilon_n)/\varepsilon_n(\gamma - \gamma'') = o(1) \quad \text{as } n \to \infty.$$

According to [4], $v(x,t) + \dfrac{k}{2t}\{\zeta(t) - x\}^2$ is a convex function of x. Thus

$$v_x(\zeta(t_0 - \varepsilon_n), t_0 - \varepsilon_n) \le v_x(\tilde{x}, t_0 - \varepsilon_n) + k\{\zeta(t_0 - \varepsilon_n) - \tilde{x}\}/(t_0 - \varepsilon_n) = o(1)$$

as $n \to \infty$ so that

$$\limsup_{t \uparrow t_0} v_x(\zeta(t), t) \ge 0.$$

However, this leads to a contradiction, since by (3.1) and (3.5) we must have

$$v_x(\zeta(t), t) = -D^+\zeta(t) \le -e^{K(\tau - t)} D^+\zeta(\tau) < -e^{K(\tau - t_0)} D^+\zeta(\tau) < 0$$

for all $t \in (\tau, t_0)$. Therefore (3.9) holds. Since $\gamma'' < \gamma$ is arbitrary, we conclude that

$$\liminf_{t \uparrow t_0} \frac{\zeta(t_0) - \zeta(t)}{t_0 - t} \ge \gamma.$$

Together with (3.8), this shows that for each $t_0 \in (\tau, T)$,

$$D^-\zeta(t_0) = \gamma = D^+\zeta(t_0). \qquad \square$$

Proof of Theorem C. By Theorem B we know that $\zeta \in C^1(t_m/\alpha, T)$. On the other hand, $\zeta \in C^1(0, t_m/\alpha)$ since $\zeta \equiv 0$ on $[0, t_m/\alpha]$. Moreover, $D^-\zeta(t_m/\alpha) = 0$. Therefore, it suffices to prove that $D^+\zeta(t_m/\alpha) = 0$.

Set $\eta = t_m/\alpha$ and assume for contradiction that

$$D^+\zeta(\eta) = \lim_{x \uparrow 0} -v_x(x, \eta) \equiv \gamma > 0. \tag{3.10}$$

Fix $t \in (0, \eta)$. By Proposition 1.2 and Taylor's theorem,

$$t_m x^2/(\eta - t) + o(x^2) = v(x,t) = v(0,t) + x v_x(0,t) + x^2 v_{xx}(\tilde{x}, t)/2,$$

where $x < \tilde{x} < 0$. Since $v(0,t) = v_x(0,t) = 0$, it follows that

$$v_{xx}(\tilde{x}, t) = 2t_m/(\eta - t) + o(1) \quad \text{as } x \uparrow 0.$$

By hypothesis, v_{xx} is a nondecreasing function of x in $(-\delta, 0) \times [0, \eta)$. Therefore

$$v_{xx}(x,t) \le 2t_m/(\eta - t) \quad \text{in } (-\delta, 0) \times (0, \eta) \tag{3.11}$$

and

$$v_{xx}(x,t) \uparrow 2t_m/(\eta - t) \quad \text{as } x \uparrow 0$$

for each $t \in (0, \eta)$.

For $\varepsilon \in (0, \eta)$ set $l \equiv \theta\varepsilon$, where

$$0 < \theta < \min(\gamma, \gamma/4t_m).$$

By Proposition 2.3,

$$v(x, \eta - \varepsilon) = L_\gamma(x, \varepsilon) + o(\varepsilon) \quad \text{for } x \in [h - \varepsilon, h + \varepsilon]$$

where $h = -\gamma\varepsilon$. Since $L_\gamma(h, \varepsilon) = 0$, it follows from the theorem of the mean that

$$-\gamma l + o(\varepsilon) = v(h, \eta - \varepsilon) - v(h - l, \eta - \varepsilon) = v_x(x', \eta + \varepsilon) l \tag{3.12}$$

and

$$o(\varepsilon) = v(h + l, \eta - \varepsilon) - v(h, \eta - \varepsilon) = v_x(x'', \eta + \varepsilon) l \tag{3.13}$$

where $h-l<x'<x''<h+l$. Define

$$\Theta_\varepsilon = \max_{|x-h|\le l} v_x(x,\eta-\varepsilon) - \min_{|x-h|\le l} v_x(x,\eta-\varepsilon).$$

Then, in view of (3.12) and (3.13),

$$\Theta_\varepsilon \ge \gamma + o(1) \quad \text{as } \varepsilon \downarrow 0.$$

Since $\theta<\gamma$ we have $h+l=(\theta-\gamma)\varepsilon<0$ and $v_{xx}(\cdot,\eta-\varepsilon)$ is continuous on $[h-l,h+l]$. There exist x_1 and x_2 such that $h-l\le x_1<x_2\le h+l$ and

$$\Theta_\varepsilon = \left| \int_{x_1}^{x_2} v_{xx}(x,\eta-\varepsilon)\,dx \right|.$$

Since $x_2-x_1\le 2l$, it follows that

$$\left| \frac{1}{x_2-x_1} \int_{x_1}^{x_2} v_{xx}\,dx \right| = \frac{1}{x_2-x_1}\Theta_\varepsilon \ge \frac{1}{2\theta\varepsilon}\{\gamma+o(1)\}.$$

On the other hand, according to (3.11), if $(\gamma+\theta)\varepsilon<\delta$, then

$$\left| \frac{1}{x_2-x_1} \int_{x_1}^{x_2} v_{xx}(x,\eta-\varepsilon)\,dx \right| \le 2t_m/\varepsilon.$$

Thus for all sufficiently small ε we have

$$2t_m \ge \frac{\{\gamma+o(1)\}}{2\theta}.$$

However, since $\theta<\gamma/4t_m$, this leads to a constradiction and we conclude that (3.10) cannot hold. On the other hand, $D^+\zeta(\eta)\ge 0$, so consequently $D^+\zeta(\eta)=0$. □

Proof of Corollary C.1. Since $v_0\le\bar{v}_0$ in $\mathbb{R}$, it follows that $v\le\bar{v}$ in $\mathbb{R}\times[0,T)$. Thus, in particular,

$$0\le\zeta(t)\le\bar{\zeta}(t) \quad \text{for } t\in[0,T), \tag{3.14}$$

where $x=\bar{\zeta}(t)$ is the right-hand interface for $\bar{v}$. By hypothesis, $t^*=\bar{t}^*=t_m/\alpha$ and $D^+\bar{\zeta}(t_m/\alpha)=0$. It therefore follows from (3.14) that $D^+\zeta(t_m/\alpha)=0$. □

Proof of Corollary C.2. Since v_0 has compact support, v exists on $\mathbb{R}\times\mathbb{R}^+$. By Corollary A.1, we have $t^*=t_m/\alpha$. Suppose that $\operatorname{supp} v_0=[-M,0]$. Let φ be a nonnegative even $C^\infty(\mathbb{R})$ function such that $\operatorname{supp}\varphi=[-M,M]$, $\varphi^{(n)}(\pm M)=0$ for all $n\ge 0$, and

$$\int_{\mathbb{R}} \varphi = 1.$$

Define $\bar{v}_0$ on $[-4M,0]$ by

$$\bar{v}_0(x) = \begin{cases} \alpha x^2 & \text{for } -M\le x<0, \\ \alpha(x+2M)^2+4\alpha M\{I_2(M)-I_2(x+2M)\} & \text{for } -2M\le x<-M \end{cases}$$

and

$$\bar{v}_0(-2M-h)=\bar{v}_0(-2M+h) \quad \text{for } 0\le h\le 2M,$$

where

$$I_1(x)\equiv\int_0^x \varphi \quad \text{and} \quad I_2(x)\equiv\int_0^x I_1.$$

It is not difficult to verify that $\bar{v}_0 \in C^\infty[-4M, 0]$, $\bar{v}_0^{(2j-1)} = 0$ at $x = 0$, $-2M$, and $-4M$ for all integers $j \geq 1$,

$$\bar{v}_0' < 0 \quad \text{on } (-2M, 0), \qquad \bar{v}_0' > 0 \quad \text{on } (-4M, -2M),$$
$$\bar{v}_0''' \geq 0 \quad \text{on } (-2M, 0), \qquad \bar{v}_0''' \leq 0 \quad \text{on } (-4M, -2M).$$

Let $\bar{v}$ denote this solution of problem (0.4) with initial datum $\bar{v}_0$. Clearly $v_0 \leq \bar{v}_0$. Moreover, it follows from the calculations in [3] that $\bar{t}^* = t_m/\alpha$ and $\bar{v}_{xxx} \geq 0$ in $(-2M, 0) \times [0, t_m/\alpha)$. In particular, $\bar{v}_{xx}$ is nondecreasing in $(-2M, 0) \times [0, t_m/\alpha)$ and the assertion follows from Corollary C.1. □

REFERENCES

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

[2] ———, *Regularity properties of flows through porous media: The interface*, Arch. Rational Mech. Anal., 37 (1970), pp. 1–10.

[3] ———, *Regularity properties of flows through porous media: A counterexample*, SIAM J. Appl. Math., 19 (1970), pp. 299–307.

[4] D. G. ARONSON AND PH. BENILAN, *Régularité des solutions de l'équation des milieux poreux dans* $\mathbb{R}^n$, C. R. Acad. Sci. Paris, 288 (1979), pp. 103–105.

[5] D. G. ARONSON AND L. A. CAFFARELLI, *The initial trace of a solution of the porous medium equation*, Trans. Amer. Math. Soc., to appear.

[6] D. G. ARONSON AND J. B. SERRIN, *Local behavior of solutions of quasilinear parabolic differential equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 81–123.

[7] G. I. BARENBLATT, *On some unsteady motions of a liquid or a gas in a porous medium*, Prikl. Math. Meh., 16 (1952), pp. 67–78. (In Russian.)

[8] L. A. CAFFARELLI AND A. FRIEDMAN, *Regularity of the free boundary for the one-dimensional flow of gas in a porous medium*, Amer. J. Math., 101 (1979), pp. 1193–1218.

[9] E. DIBENEDETTO, *Regularity results for the porous medium equation*, Ann. Mat. Pura Appl., 121 (1981), pp. 249–262.

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

[11] A. S. KALASHNIKOV, *The Cauchy problem in the class of increasing functions for equations of the nonstationary seepage type*, Vestnik Moskov. Univ. Ser. 1 Mat. Meh., 6 (1963), pp. 17–27. (In Russian.)

[12] ———, *On the occurrence of singularities in the solutions of the equation of nonstationary filtration*, Z. Vyčisl. Mat. i Mat. Fiz., 7 (1967), pp. 440–444. (In Russian.)

[13] B. F. KNERR, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.

[14] S. N. KRUZHKOV, *Results concerning the nature of the continuity of solutions of parabolic equations and some of their applications*, Math. Notes, 6 (1969), pp. 517–523.

[15] A. A. LACEY, J. R. OCKENDON AND A. B. TAYLER, *'Waiting-time' solutions of a nonlinear diffusion equation*, SIAM J. Appl. Math., 42 (1982), pp. 1252–1264.

[16] O. A. OLEINIK, A. S. KALASHNIKOV AND CHZOU YUI-LIN, *The Cauchy problem and boundary problems for equations of the type of unsteady filtration*, Izv. Akad. Nauk SSSR Ser. Mat., 22 (1958), pp. 667–704. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0002 $01.25/0

DECAY TO UNIFORM STATES IN COMPETITIVE SYSTEMS*

PETER N. BROWN†

Abstract. Weakly coupled parabolic systems describing populations undergoing diffusion and competition in a spatial domain Ω are discussed. Assuming the existence of a unique critical point of the interaction dynamics with all populations coexisting, sufficient conditions are given (on the dynamics) which guarantee the global asymptotic stability of the critical point. These conditions imply the existence of a continuous family of contracting rectangles which decrease down to the critical point and then the theory of contracting rectangles developed in [2] implies the stability. General two and n species models are discussed along with some illustrative examples.

1. Introduction.

We consider systems of parabolic equations whose general form is

$$\partial_t u_i = d_i \Delta u_i + u_i M_i(u_1, \cdots, u_n) \quad \text{in } \Omega \times \mathbb{R}^+, \tag{1.1}$$

$$u_i(x,0) = \phi_i(x) \quad \text{for } x \text{ in } \Omega, \tag{1.2}$$

$$\frac{\partial u_i}{\partial n} \equiv 0 \quad \text{on } \partial\Omega \times \mathbb{R}^+, \tag{1.3}$$

where $i = 1, \cdots, n$, Ω is a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary, the constants $d_1, \cdots, d_n$ are all positive, the M_i are all smooth functions of $u = (u_1, \cdots, u_n)$ and $\partial u_i / \partial n$ is the normal derivative in the direction of the outward normal at a point $x \in \partial\Omega$.

System (1.1)–(1.3) is an example of a system of *reaction-diffusion* equations, and describes the growth of n populations which are both diffusing and interacting in Ω. The function $u_i(x,t)$ $(i = 1, \cdots, n)$ represents the ith population density, and the boundary conditions (1.3) have the effect of confining the populations to the spatial habitat Ω, i.e., there is no migration across the boundary of Ω.

We study here the following problem: Given that the vector field $(u_1 M_1, \cdots, u_n M_n)$ has exactly one critical point u^* in the positive orthant, what extra assumptions on $M_1, \cdots, M_n$ will guarantee that for all solutions $u(x,t)$ of (1.1)–(1.3) with $\phi_i(x) \geq 0$ and $\phi_i(x) \not\equiv 0$ for x in $\Omega (i = 1, \cdots, n)$, we have that

$$\lim_{t \to \infty} u(x,t) = u^*,$$

i.e., what conditions on the M_i will guarantee the global stability of u^*? For example, when $n = 1$, (1.1) reduces to

$$u_t = d\Delta u + uM(u).$$

Then, assuming there exists a critical point $u^* > 0$, it is easily seen that a sufficient condition guaranteeing the global stability of u^* is just

$$(u - u^*) M(u) < 0 \quad \text{for all } u > 0,\ u \neq u^*.$$

We will restrict attention here to competitive systems only, i.e., those for which

$$\frac{\partial M_i}{\partial u_j} < 0 \quad \text{for } i \neq j,$$

so that a growth in the jth population is harmful to the ith population. In [2] we considered competition models of Lotka-Volterra type, where the functions $M_1, \cdots, M_n$

* Received by the editors January 21, 1982.
† Department of Mathematics, University of Houston, Houston, Texas 77004.

all depend linearly on the components of u. There, the conditions assumed actually are sufficient to imply the existence, uniqueness and global stability of a critical point u^*. In general, however, the existence and uniqueness of u^* must be postulated along with some additional assumptions which generalize those given in [2]. Recent other work on global stability in many population (or species) systems includes that of Albrecht et al [1], Goh [6], Hastings [7], [8] and DeMottoni and Rothe [5].

In §2 we begin with some preliminaries and then state some results on invariant and contracting rectangles. In §3 we consider a general two-species competition model and then show that, basically, the condition implying the global stability of a critical point in the Lotka-Volterra competition model (given in [2]) also suffices for the more general case. Finally, in §4 we consider a general n-species model. We give conditions guaranteeing the global stability of a critical point and then consider some illustrative examples.

2. Preliminaries. In this section we give some background material on contracting rectangles. We consider systems of the form

$$u_t = D\Delta u + F(u) \quad \text{in } \Omega\times\mathbb{R}^+, \tag{2.1}$$

$$u(x,0) = u^0(x) \quad \text{for } x \text{ in } \Omega, \tag{2.2}$$

$$\frac{\partial u}{\partial n} \equiv 0 \quad \text{on } \partial\Omega\times\mathbb{R}^+, \tag{2.3}$$

where $u=(u_1,\cdots,u_n)$, $F:\mathbb{R}^n\to\mathbb{R}^n$ is smooth, $D=\operatorname{diag}\{d_1,\cdots,d_n\}$, $d_i>0$ for all i and Ω is a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary.

DEFINITION 2.1. An n-dimensional rectangle $\Sigma=\{u: a\le u\le b\}$ (where $a\le u\le b$ means $a_i\le u_i\le b_i$ for $i=1,\cdots,n$) with $-\infty\le a_j\le b_j\le+\infty$ is said to be *invariant* for (2.1)–(2.3) if when $u^0(x)\in\Sigma$ for all x in Ω, it follows that $u(x,t)\in\Sigma$ for all $t>0$ and x in Ω.

A necessary and sufficient condition for Σ to be invariant (see [3] and [4]) is that the vector field $F=(F_1,\cdots,F_n)$ does not point out of Σ, i.e., for $u\in\Sigma$

$$F_i(u)\ge 0 \quad \text{when } u_i=a_i \tag{2.4}$$

and

$$F_i(u)\le 0 \quad \text{when } u_i=b_i. \tag{2.5}$$

DEFINITION 2.2. $\Sigma=\{u: a\le u\le b\}$ is said to be *contracting* for the vector field F if for each i, $F_i>0$ in (2.4) and $F_i<0$ in (2.5). Note that Σ contracting implies that Σ is invariant.

DEFINITION 2.3. Let u^* be an isolated critical point of the vector field F, i.e., $F(u^*)=0$. Let $\Sigma(\tau)=\{u: a(\tau)\le u\le A(\tau)\}$, defined for $0\le\tau\le 1$, be a one-parameter family of rectangles. Then $\Sigma(\tau)$ is said to be a *decreasing family of contracting rectangles for* u^* if

(i) $\Sigma(1)=\{u^*\}$,

(ii) $a(\tau)$ and $A(\tau)$ are continuous with $a(\tau)$ increasing and $A(\tau)$ decreasing,

(iii) $\Sigma(\tau)$ contracting for $0\le\tau<1$.

Next, let $\Sigma=\{u: a\le u\le b\}$ be an invariant rectangle for (2.1)–(2.3). Associated with Σ are the maximal and minimal functions

$$F_i^+(u)=\max\{F_i(\theta_1,\cdots,\theta_{i-1},u_i,\theta_{i+1},\cdots,\theta_n): a_j\le\theta_j\le u_j, j\ne i\}$$

and

$$F_i^-(u)=\min\{F_i(\theta_1,\cdots,\theta_{i-1},u_i,\theta_{i+1},\cdots,\theta_n): u_j\le\theta_j\le b_j, j\ne i\}$$

for $i=1,\cdots,n$. In [4] it is shown that $F_1^{\pm},\cdots,F_n^{\pm}$ are all locally Lipschitz continuous in Σ so that the ordinary differential equations

$$\frac{du}{dt}=F^{+}(u), \qquad u(0)=w^0 \tag{2.6}$$

and

$$\frac{du}{dt}=F^{-}(u), \qquad u(0)=v^0 \tag{2.7}$$

have unique solutions. The following result is a special case of the comparison theorem in [4]:

LEMMA 2.4. *Let Ω be a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary. Let F be smooth and let $u(x,t)$ be the solution of* (2.1)–(2.3). *Let $u^{+}(t)$ and $u^{-}(t)$ be the respective solutions of* (2.6) *and* (2.7).

Then $v^0\le u^0(x)\le w^0$ for all x in Ω implies that $u^{-}(t)\le u(x,t)\le u^{+}(t)$ for all x in Ω and $t\ge 0$.

In [2] the following local asymptotic stability result is proved:

THEOREM 2.5. *Let Ω be a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary. Let F be smooth and assume there is an isolated critical point u^* of F. Let $\Sigma(\tau)$ be a decreasing family of contracting rectangles for u^*.*

Then if $u(x,t)$ is a solution of (2.1)–(2.3) *satisfying, for some τ in $[0,1)$, $u(x,T)\in\Sigma(\tau)$ for all x in Ω and some $T\ge 0$, then*

$$\lim_{t\to\infty} u(x,t)=u^*, \quad \textit{uniformly for } x \textit{ in } \Omega.$$

We finish this section with some notation and a final definition. Let

$$R^0=\{(u,v): u>0, v>0\}, \qquad R=\overline{R^0},$$

and

$$R_n^0=\{u: u_i>0 \text{ for } i=1,\cdots,n\}, \qquad R_n=\overline{R_n^0}.$$

DEFINITION 2.6. Let $F(u)=(u_1M_1,\cdots,u_nM_n)$ and let $F(u^*)=0$. Then u^* is said to be a *feasible equilibrium* of F if $u_i^*>0$ for $i=1,\cdots,n$.

3. Two-species competition. In this section we consider a general model for the interaction of two competing species, with diffusion effects included. Let $u(x,t)$ and $v(x,t)$ be the solutions of the initial-boundary value problem

$$\begin{aligned} u_t&=d_1\Delta u+uM(u,v),\\ v_t&=d_2\Delta v+vN(u,v), \end{aligned} \qquad \text{in } \Omega\times\mathbb{R}^{+}, \tag{3.1}$$

$$\begin{aligned} u(x,0)&=u^0(x),\\ v(x,0)&=v^0(x), \end{aligned} \qquad \text{for } x \text{ in } \Omega, \tag{3.2}$$

$$\frac{\partial u}{\partial n}\equiv 0\equiv\frac{\partial v}{\partial n} \quad \text{for } (x,t)\in\partial\Omega\times\mathbb{R}^{+}. \tag{3.3}$$

Here Ω is a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary, and the diffusion coefficients d_1 and d_2 are positive constants.

We will make the following assumptions on the functions M and N (see [1], [8] and [9] for ecological interpretations of these conditions):

(3.4) M and N are both C^1 in R with M_u, M_v, N_u, $N_v<0$ in R^0 (where $M_u=\partial M/\partial u$, etc.).

(3.5) There exist positive constants A and B such that
(a) $(u-A)M(u,0)<0$ whenever $u\geq 0$, $u\neq A$ and
(b) $(v-B)N(0,v)<0$ whenever $v\geq 0$, $v\neq B$.

(3.6) There exist positive constants C and D such that
(a) $(v-D)M(0,v)<0$ whenever $v\geq 0$, $v\neq D$ and
(b) $(u-C)N(u,0)<0$ whenever $u\geq 0$, $u\neq C$.

(3.7) There exists a unique feasible equilibrium (u^*,v^*).

(3.8) The constants A,B,C and D are such that $A<C$ and $B<D$.

We begin with two lemmas.

LEMMA 3.1. *Assume that* (3.4)–(3.6) *hold. Then there exist continuous functions* $k_1(v)$ *and* $k_2(u)$, *defined on* $0\leq v\leq D$ *and* $0\leq u\leq C$, *respectively, such that*

(3.9) (a) $k_1(0)=A$, $k_1(D)=0$, *and* $0<k_1(v)<A$ *for* $0<v<D$,
(b) k_1 *is* C^1 *on* $(0,D)$ *with* $k_1'<0$ *there*,
(c) $u>k_1(v)$ *iff* $M(u,v)<0$ *and* $u=k_1(v)$ *iff* $M(u,v)=0$ *in the rectangle* $Q_1\equiv\{(u,v): 0\leq u\leq A,\ 0\leq v\leq D\}$

and

(3.10) (a) $k_2(0)=B$, $k_2(C)=0$, *and* $0<k_2(u)<B$ *for* $0<u<C$,
(b) k_2 *is* C^1 *on* $(0,C)$ *with* $k_2'<0$ *there*,
(c) $v>k_2(u)$ *iff* $N(u,v)<0$ *and* $v=k_2(u)$ *iff* $N(u,v)=0$ *in the rectangle* $Q_2\equiv\{(u,v): 0\leq u\leq C,\ 0\leq v\leq B\}$.

Proof. We prove only the results involving M since those concerned with N may be shown analogously.

First, note that (3.4), (3.5a) and (3.6a) imply that $M<0$ in $R\setminus Q_1$ (see Fig. 1).

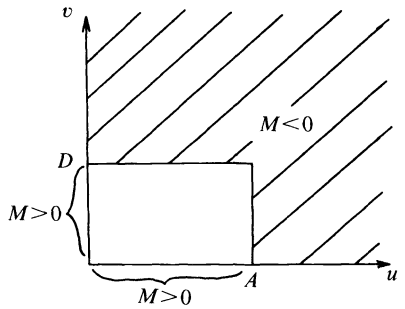


FIG. 1

Next, $M_u<0$ implies that for $\alpha>0$,

(3.11) $M(u,\alpha)$ is a strictly decreasing function for $0\leq u<\infty$.

Now, (3.4), (3.11), the fact that $M(0,v)>0$ for $0\leq v<D$, that $M(A,v)<0$ for $0<v\leq D$, $M(0,D)=0$ and $M(A,0)=0$ all imply that for each v in $[0,D]$ there exists a $u=k_1(v)$ such that

$$M(k_1(v),v)=0.$$

The function k_1 is defined on $[0,D]$ with $k_1(D)=0$ and $k_1(0)=A$. The implicit function theorem then implies k_1 is differentiable at any point v in $(0,D)$ and that

$$k_1'(v)=-\frac{M_v(k_1(v),v)}{M_u(k_1(v),v)}<0.$$

Hence, k_1 is a continuous decreasing function on $(0,D)$ with $0\le k_1(v)\le A$ for $v\in[0,D]$. We show that k_1 is continuous on $[0,D]$. Since k_1 is continuous, decreasing and bounded on $(0,D)$, the limits

$$a=\lim_{v\to 0^+} k_1(v) \quad \text{and} \quad b=\lim_{v\to D^-} k_1(v)$$

both exist. Suppose that $b>0$. Then by the continuity of M, $M(b,D)=0$. But $M(0,D)=0$ by (3.6a), which contradicts $M_u<0$ in $\mathbb{R}^0$. Next, suppose $a<A$. Again, by the continuity of M, $M(a,0)=0$, which leads to a contradiction of (3.5a). Therefore, we have shown (3.9a) and (3.9b). Since (3.9c) follows easily, this completes the proof of the lemma. □

LEMMA 3.2. *Assume that* (3.4)–(3.8) *hold. Then there exists a decreasing family of rectangles* $\Sigma(\tau)$, *defined for* $0\le\tau\le 1$, *such that, when* $0<\tau<1$, $\Sigma(\tau)$ *is contracting for the vector field* (uM,vN)

$$\Sigma(1)=\{(u^*,v^*)\} \tag{3.12}$$

and

$$\Sigma(0)=\{(u,v):(0,0)\le(u,v)\le\tfrac{1}{2}(A+C,B+D)\}. \tag{3.13}$$

Proof. Since the conditions of Lemma 3.1 hold, there exist functions $k_1(v)$ and $k_2(u)$ satisfying (3.9) and (3.10). From the proof of that lemma we also have that $N<0$ in $R\setminus Q_2$ and $M<0$ in $R\setminus Q_1$. By (3.7) and (3.8) we have that the graphs of k_1 and k_2 only intersect at the point (u^*,v^*) (i.e., $u^*=k_1(v^*)$ and $v^*=k_2(u^*)$). Furthermore,

$$0<u^*<A \quad \text{and} \quad 0<v^*<B. \tag{3.14}$$

Therefore, the phase portrait of the vector field (uM,vN) has the qualitative form

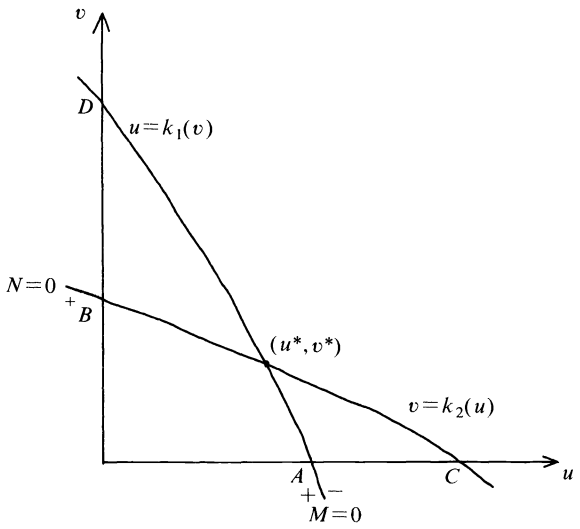


FIG. 2

with $k_1(k_2(0))>0$ and $k_2(k_1(0))>0$. Note that these last two inequalities are equivalent to (3.8).

We next show that

$$k_1(k_2(\tau u^*))>\tau u^* \tag{3.15}$$

and

$$k_2(k_1(\tau v^*))>\tau v^*, \tag{3.16}$$

whenever $0\leq\tau<1$.

Suppose there exists a value τ_1 in $(0,1)$ for which $k_1(k_2(\tau_1 u^*))=\tau_1 u^*$. Then by (3.9c)

$$M(\tau_1 u^*, k_2(\tau_1 u^*))=0$$

and by (3.10c)

$$N(\tau_1 u^*, k_2(\tau_1 u^*))=0.$$

Since $0<\tau_1<1$ we have $0<\tau_1 u^*<u^*$ and hence

$$k_2(\tau_1 u^*)>k_2(u^*)=v^*>0$$

by (3.10b). Therefore, $(\tau_1 u^*, k_2(\tau_1 u^*))$ is a feasible equilibrium and hence must equal (u^*, v^*) by (3.7) which implies $\tau_1=1$. Since this is a contradiction we have (3.15). Condition (3.16) can be proved similarly.

To construct $\Sigma(\tau)=\{(u,v): (a(\tau),b(\tau))\leq(u,v)\leq(A(\tau),B(\tau))\}$ let $a(\tau)=\tau u^*$, $b(\tau)=\tau v^*$,

$$A(\tau)=\frac{k_1(\tau v^*)+k_2^{-1}(\tau v^*)}{2}$$

and

$$B(\tau)=\frac{k_2(\tau u^*)+k_1^{-1}(\tau u^*)}{2}$$

for $0\leq\tau\leq1$, where k_1^{-1} and k_2^{-1} are the respective inverses of k_1 and k_2, which exist by (3.9) and (3.10). Note that (3.15) and (3.16) imply

$$k_2(\tau u^*)<k_1^{-1}(\tau u^*) \quad \text{and} \quad k_1(\tau v^*)<k_2^{-1}(\tau v^*) \quad \text{for } 0\leq\tau<1.$$

Hence,

$$u^*<k_1(\tau v^*)<A(\tau)<k_2^{-1}(\tau v^*)\leq A \tag{3.17}$$

and

$$v^*<k_2(\tau u^*)<B(\tau)<k_1^{-1}(\tau u^*)\leq B \tag{3.18}$$

for $0\leq\tau<1$ with $A(1)=u^*$ and $B(1)=v^*$. Thus, $\Sigma(\tau)$ is decreasing,

$$\Sigma(0)=\left\{(u,v): (0,0)\leq(u,v)\leq\left(\frac{A+C}{2},\frac{B+D}{2}\right)\right\}$$

and

$$\Sigma(1)=\{(u^*,v^*)\}.$$

It remains to show $\Sigma(\tau)$ is contracting for $0<\tau<1$. Let $\tau\in(0,1)$ and consider $M(a(\tau),v)$ for $b(\tau)\le v\le B(\tau)$. By (3.4), (3.9), and (3.18)

$$M(a(\tau),v)\ge M(a(\tau),B(\tau))>M\big(a(\tau),k_1^{-1}(\tau u^*)\big)$$
$$=M\big(\tau u^*,k_1^{-1}(\tau u^*)\big)=0.$$

Next, consider $M(A(\tau),v)$ for $b(\tau)\le v\le B(\tau)$. By (3.4), (3.9) and (3.17)

$$M(A(\tau),v)\le M(A(\tau),b(\tau))<M\big(k_1(\tau v^*),b(\tau)\big)$$
$$=M\big(k_1(\tau v^*),\tau v^*\big)=0.$$

Similarly, one can show

$$N(u,b(\tau))>0 \quad \text{and} \quad N(u,B(\tau))<0$$

for $a(\tau)\le u\le A(\tau)$. Therefore, for each $\tau\in(0,1)$, $\Sigma(\tau)$ is a contracting rectangle for the vector field (uM,vN), since $a(\tau)=\tau u^*>0$ and $b(\tau)=\tau v^*>0$. □

We can now prove the main result of this section.

THEOREM 3.3. *Let Ω be a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary, and let d_1 and d_2 be any positive constants. Let conditions* (3.4)–(3.8) *hold, and let $(u(x,t),v(x,t))$ be a solution of* (3.1)–(3.3) *with bounded nonnegative initial conditions $u^0(x)$ and $v^0(x)$ satisfying*

$$u^0(x)\not\equiv 0 \quad \textit{and} \quad v^0(x)\not\equiv 0 \quad \textit{for } x \textit{ in } \Omega. \tag{3.19}$$

Then

$$\lim_{t\to\infty}(u(x,t),v(x,t))=(u^*,v^*) \quad \textit{uniformly for } x \textit{ in } \Omega. \tag{3.20}$$

Proof. Since $u^0(x)$ and $v^0(x)$ are bounded, there exist constants $A_0>C$ and $B_0>D$ such that the rectangle

$$\Sigma_0=\{(u,v):(0,0)\le(u,v)\le(A_0,B_0)\}$$

is invariant for (3.1)–(3.3) and

$$\big(u^0(x),v^0(x)\big)\in\Sigma_0 \quad \text{for all } x \text{ in } \Omega.$$

Let $(u^+(t),v^+(t))$ be the solution of the maximal ODE (2.6)–(2.7) corresponding to Σ_0. Then by Lemma 2.4

$$u(x,t)\le u^+(t) \quad \text{and} \quad v(x,t)\le v^+(t)$$

for x in Ω, $t\ge 0$ and

$$u^+(t)\searrow A \quad \text{and} \quad v^+(t)\searrow B$$

as $t\nearrow\infty$. Hence, there exists a time $t_1>0$ such that

$$A<u^+(t_1)<\frac{A+C}{2}$$

and

$$B<v^+(t_1)<\frac{B+D}{2}.$$

Now, Ω bounded implies $\bar{\Omega}$ compact. Then, since (3.19) holds, the strong maximum principle (cf. [4, Thm. 2]) implies

$$u(x,t_1)>\delta>0 \quad \text{and} \quad v(x,t_1)>\varepsilon>0 \quad \text{for } x \text{ in } \bar{\Omega}$$

for some constants δ and ε. Therefore, by (3.13), there exists a τ_1 in $(0,1)$ such that

$$(u,(x,t_1),v(x,t_1))\in\Sigma(\tau_1) \quad \text{for all } x \text{ in } \bar{\Omega},$$

where $\Sigma(\tau)$, $0\leq\tau\leq 1$, is the family of rectangles given by Lemma 3.2. Thus, by Theorem 2.5, (3.20) holds. □

We note that condition (3.8) is essential in the proof of Theorem 3.3. The solution $(u^+(t),v^+(t))$ of the maximal ODE's (2.6) and (2.7) must enter $\Sigma(0)$ in finite time. To see that (3.8) is also necessary, in some sense, we consider the following example:

Example 3.4. Let

$$M(u,v)=k(u)-v-\varepsilon v(u-1)$$

and

$$N(u,v)=-5(u-1)-v,$$

where $0<\varepsilon<1$ and

$$k(u)=\begin{cases}-10(u-1)^3, & 0\leq u\leq 1,\\ -(u-1)^3, & u>1 \quad .\end{cases}$$

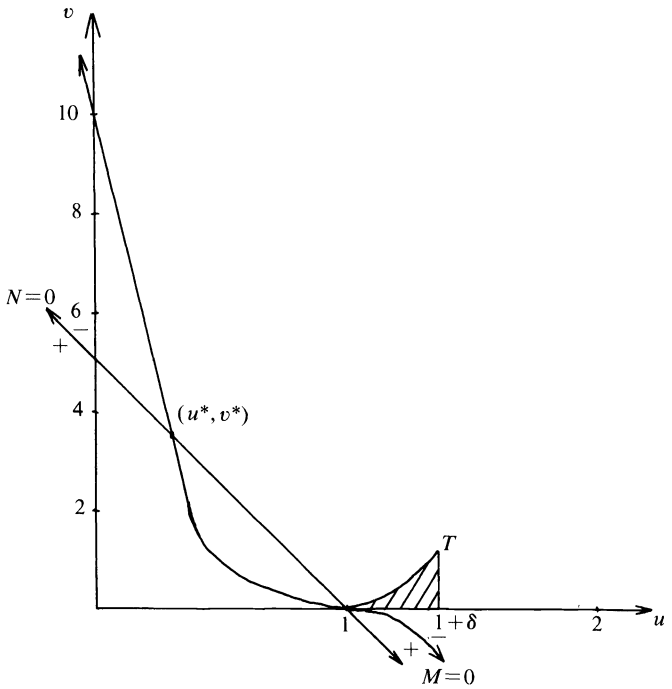


FIG. 3

Conditions (3.4)–(3.7) are readily verified with

$$A=C=1, \quad B=5, \quad D=\frac{10}{1-\varepsilon}$$

and (u^*,v^*) is given by

$$u^*=1+\frac{1}{4}\left(\varepsilon-\sqrt{\varepsilon^2+8}\right) \quad \text{and} \quad v^*=-5(u^*-1).$$

Let $T=\{(u,v): 1\le u\le 1+\delta,\ 0\le v\le (u-1)^2\}$, $\delta>0$. For δ small enough, it is easy to check that the vector field $F(u,v)=(uM,vN)$ does not point out of T, so that T is invariant for the ODE

$$\frac{du}{dt}=uM(u,v), \qquad \frac{dv}{dt}=vN(u,v).$$

Any trajectory entering T must therefore stay there and converge to the critical point $(1,0)$ as $t\to\infty$. Thus, (u^*,v^*) is no longer globally stable, even for the ODE. However, (u^*,v^*) is locally stable (see Fig. 3).

4. *n*-species competition. In this section we consider a general model for the competitive interaction of n-species. Let $u(x,t)=(u_1(x,t),\cdots,u_n(x,t))$ be the solution of the initial-boundary value problem

$$\partial_t u_i=d_i\Delta u_i+u_iM_i(u_1,\cdots,u_n) \quad \text{for } (x,t) \text{ in } \Omega\times\mathbb{R}^+, \tag{4.1}$$

$$u_i(x,0)=u_i^0(x) \quad \text{for } x \text{ in } \Omega, \tag{4.2}$$

$$\frac{\partial u_i}{\partial n}\equiv 0 \quad \text{for } (x,t)\in\partial\Omega\times\mathbb{R}^+ \tag{4.3}$$

for $i=1,\cdots,n$. Here Ω is a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary, and the diffusion coefficients $d_1,\cdots,d_n$ are all positive constants.

We will make the following assumptions on the functions $M_1(u),\cdots,M_n(u)$, $u=(u_1,\cdots,u_n)$:

(4.4) (a) $M_1,\cdots,M_n$ are all C^1 in a neighborhood of R_n,
(b) $\partial M_i/\partial u_j<0$ in R_n^0 for $i\neq j$ $(i,j=1,\cdots,n)$,
(c) $\partial M_i/\partial u_i<0$ in a neighborhood of R_n $(i=1,\cdots,n)$.

(4.5) There exist positive constants A_{ij} such that

$$(u_i-A_{ij})M_j(0,\cdots,0,u_i,0,\cdots,0)<0 \quad \text{for } u_i\ge 0,\ u_i\neq A_{ij},\ i,j=1,\cdots,n$$

with u_i in the ith place in $(0,\cdots,0,u_i,0,\cdots,0)$,

(4.6) There exists a unique feasible equilibrium $u^*=(u_1^*,\cdots,u_n^*)$ of the vector field $(u_1M_1,\cdots,u_nM_n)$.

(4.7) The constants A_{ij} satisfy $A_{ii}<A_{ij}$ for $i\neq j$, $i,j=1,\cdots,n$.

One can show as in the two-dimensional case the following lemma:

LEMMA 4.1. *Assume* (4.4) *and* (4.5) *hold. Then there exist* C^1 *functions* $k_1,\cdots,k_n$ *such that for each* i

(4.8) (a) $M_i(u)\underset{(<)}{=}0$ *iff* $u_i\underset{(>)}{=}k_i(\hat{u})$, *in the rectangle* $\{0\le u\le(A_{1i},\cdots,A_{ni})\}$ *where* $\hat{u}=(u_1,\cdots,u_{i-1},u_{i+1},\cdots,u_n)$,
(b) $\partial k_i/\partial u_j(\hat{u})<0$ for $j\neq i$, $\hat{u}\ge 0$ *and in the domain of* k_i,

(c) $0 \le k_i(\hat{u}) \le A_{ii}$, $k_i(\hat{0}) = A_{ii}$ *and* $k_i(0,\cdots,0,A_{ji},0,\cdots,0) = 0$ *for* $j \ne i$,

(d) $\{\hat{u} \ge 0 : k_i(\hat{u}) \ge 0$ *and* $\hat{u}$ *in the domain of* $k_i\}$ *is a closed and bounded set.*

The next lemma gives the existence of a family of contracting rectangles, but this case differs from the $n=2$ case in that now, the generalization of the conditions $k_1(k_2(0))>0$ and $k_2(k_1(0))>0$ must be assumed to hold over a whole interval, not just at a single point. (Compare (3.15) and (3.16) with (4.9).) We give an example below to illustrate this fact.

LEMMA 4.2. *Let* $V(\tau) = (k_1(\tau\hat{u}^*),\cdots,k_n(\tau\hat{u}^*))$ *for* $0 \le \tau \le 1$. *Suppose* (4.4)–(4.7) *hold, and that*

$$k_i(\hat{V}(\tau)) > \tau u_i^* \quad \text{for } 0 \le \tau < 1. \tag{4.9}$$

Then there exist an n-dimensional vector A and a continuous decreasing family of rectangles $\Sigma(\tau)$, *defined for* $0 \le \tau \le 1$, *such that*

$$\Sigma(0) = \{u : 0 \le u \le A\} \text{ with } A > (A_{11},\cdots,A_{nn}), \tag{4.10}$$

$$\Sigma(1) = \{u^*\}, \tag{4.11}$$

$$\Sigma(\tau) \text{ is contracting for } 0 < \tau < 1 \text{ for the vector field } (u_1 M_1,\cdots,u_n M_n). \tag{4.12}$$

Remark. Condition (4.9) actually involves an assumption and an implication. First, condition (4.9) implicitly assumes that $\hat{V}(0) \in$ domain of k_i $(i=1,\cdots,n)$, and since $V(\tau)$ is decreasing, this implies that $V(\tau) \in$ domain of k_i $(i=1,\cdots,n)$ for all $0 \le \tau \le 1$. Furthermore, this gives $u^* < V(\tau)$, $0 \le \tau < 1$, with $u^* = V(1)$. Second, condition (4.9) actually implies (4.7), since (4.7) not true implies $\hat{V}(0) \notin$ domain of k_i (for *some* i). At first glance, (4.7) may seem to be the natural generalization of condition (3.8). However, even in the case of linear M_i (i.e., Lotka–Volterra dynamics) it can be shown that when (4.7) holds and (4.9) does not, a critical point stable for the ODE may be unstable for the PDE, (i.e., the critical point can be destablized by the introduction of diffusion terms).

Proof. We show below the existence of continuous functions $\alpha_1(\tau),\cdots,\alpha_n(\tau)$, defined for $0 \le \tau \le 1$, such that

$$k_i(\alpha_i(\tau)\hat{V}(\tau)) = \tau u_i^*, \qquad 0 \le \tau \le 1, \quad i=1,\cdots,n, \tag{4.13}$$

and

$$\alpha_i(1) = 1, \quad \alpha_i(\tau) > 1, \quad 0 \le \tau < 1, \quad i=1,\cdots,n.$$

Assuming this for the moment, we construct the family of rectangles $\Sigma(\tau)$ and show that properties (4.10)–(4.12) hold. Let

$$\bar{\alpha}_i(\tau) = \min_{0 \le s \le \tau} \alpha_i(s) \quad \text{for } i=1,\cdots,n. \tag{4.14}$$

Then let

$$a(\tau) = \tau u^*$$

and, for $i=1,\cdots,n$,

$$A_i(\tau) = \frac{1}{2}\left(1 + \min_{j \ne i} \{\bar{\alpha}_j(\tau)\}\right) V_i(\tau). \tag{4.15}$$

Finally, define $\Sigma(\tau)$ by

$$\Sigma(\tau) = \{u : a(\tau) \le u \le A(\tau)\},$$

where $A(\tau)=(A_1(\tau),\cdots,A_n(\tau))$ with $A=(A_1(0),\cdots,A_n(0))$. Clearly, $\Sigma(\tau)$ is a continuous decreasing family of rectangles satisfying (4.10) and (4.11). It remains to show that (4.12) holds or that $\Sigma(\tau)$ is contracting for $0<\tau<1$. To show this it is enough to show that, for each $i=1,\cdots,n$ and τ in $(0,1)$,

$$M_i(A_1,\cdots,A_{i-1},a_i,A_{i+1},\cdots,A_n)>0 \tag{4.16}$$

and

$$M_i(a_1,\cdots,a_{i-1},A_i,a_{i+1},\cdots,a_n)<0. \tag{4.17}$$

Here we have suppressed the τ dependence for convenience.

To show (4.16) it is equivalent to show

$$a_i<k_i(\hat{A}). \tag{4.18}$$

But

$$a_i(\tau)=\tau u_i^*=k_i\big(\alpha_i(\tau)\hat{V}(\tau)\big)\le k_i\big(\bar{\alpha}_i(\tau)\hat{V}(\tau)\big)$$

by (4.8) and since $\bar{\alpha}_i(\tau)\le\alpha_i(\tau)$ by (4.14). We show that

$$\bar{\alpha}_i(\tau)\hat{V}(\tau)>\hat{A}(\tau) \tag{4.19}$$

which will give us (4.18) by (4.8). Since $\alpha_i(\tau)>1$ implies $\bar{\alpha}_i(\tau)>1$ we easily see that (4.19) holds by (4.15).

Next, to show the second inequality (4.17), it is equivalent to show

$$A_i>k_i(\hat{a})=k_i(\tau\hat{u}^*)=V_i(\tau).$$

Again, by (4.15) and since $\bar{\alpha}_i(\tau)>1$ we have that

$$A_i(\tau)>V_i(\tau).$$

Therefore, $\Sigma(\tau)$ is contracting for $0<\tau<1$.

Finally, we show the existence of the functions $\alpha_1(\tau),\cdots,\alpha_n(\tau)$ satisfying (4.13). For each $i=1,\cdots,n$ define

$$h_i(\alpha,\tau)=k_i\big(\alpha\hat{V}(\tau)\big)-\tau u_i^*.$$

Since the set $\{\hat{u}\ge 0: k_i(\hat{u})\ge 0\}$ is closed and bounded, and since (4.9) holds, for each τ in $[0,1]$ there exists a unique $\beta_i=\beta_i(\tau)$ such that

$$k_i\big(\beta_i(\tau)\hat{V}(\tau)\big)=0.$$

Moreover, by (4.8) and (4.9) $\beta_i(\tau)>1$ for τ in $[0,1]$ and by the implicit function theorem $\beta_i(\tau)$ is continuous in $[0,1]$. Hence, for each τ in $[0,1]$

$$h_i(0,\tau)>0 \quad\text{and}\quad h_i\big(\beta_i(\tau),\tau\big)<0.$$

Since $\partial h_i/\partial\alpha<0$, for each τ in $[0,1]$ there exists a function $\alpha_i(\tau)$ such that

$$h_i\big(\alpha_i(\tau),\tau\big)=0 \quad\text{in } 0\le\tau\le 1.$$

The implicit function theorem then implies that $\alpha_i(\tau)$ is continuous in $[0,1]$, and (4.9) gives $\alpha_i(\tau)>1$ for $0\le\tau<1$. Since we have equality in (4.9) when $\tau=1$, clearly $\alpha_i(1)=1$. This completes the proof of the lemma. □

An argument similar to that used to prove Theorem 3.3 gives the main result of this section.

THEOREM 4.3. *Let Ω be a bounded domain in $\mathbb{R}^m$ with sufficiently smooth boundary and $d_1,\cdots,d_n$ be any positive constants. Let conditions* (4.4)–(4.7) *and* (4.9) *hold, and let*

$u(x,t)$ be a solution of (4.1)–(4.3) *with bounded nonnegative initial conditions* $u^0(x)$ *satisfying*

$$u_i^0(x) \not\equiv 0 \quad \text{for } x \text{ in } \Omega \text{ and } i=1,\cdots,n. \tag{4.20}$$

Then

$$\lim_{t\to\infty} u(x,t)=u^* \quad \text{uniformly for } x \text{ in } \Omega. \tag{4.21}$$

Remarks. Condition (4.9) seems new even for the ordinary differential equation. To our knowledge, this is the first list of sufficient conditions guaranteeing the global stability of a critical point in a general n-species interaction model.

We next illustrate the use of the theorem on an example.

Example 4.4. For $i=1,\cdots,n$ let

$$M_i(u)=b-u_i-\sum_{\substack{j=1\\ j\neq i}}^{n} f(u_j), \tag{4.22}$$

where $b>0$ is a constant and $f(v)$, defined for all v in a neighborhood of $[0,\infty)$, is C^1 and satisfies

(4.23) (a) $f(0)=0$,
(b) $0<f'(v)<1/(n-1)$ for $v>0$,
(c) the equation $f(v)=b$ has a solution,
(d) $f(b)<b/(n-1)$.

The function

$$f(v)=\frac{\delta v}{1+\varepsilon v}, \tag{4.24}$$

where $0<\delta<1/(n-1)$ and $0<\varepsilon<\delta/b$, satisfies (4.23) and is also an interaction term of Holling type [10].

We prove the following result:

THEOREM 4.5. *Let Ω be a bounded region in $\mathbb{R}^m$ with sufficiently smooth boundary and $d_1,\cdots,d_n$ be any positive constants. Let $M_1,\cdots,M_n$ be defined by* (4.22) *and* (4.23).

Then there exists a unique feasible equilibrium u^* which is globally asymptotically stable in the sense that for $u(x,t)$ a solution of (4.1)–(4.3) *with bounded nonnegative initial conditions $u^0(x)$ satisfying*

$$u_i^0(x) \not\equiv 0 \quad \text{for } x \text{ in } \Omega \text{ and } i=1,\cdots,n,$$

we have

$$\lim_{t\to\infty} u(x,t)=u^* \quad \text{uniformly for } x \text{ in } \Omega.$$

Proof. We need only verify that conditions (4.4)–(4.7) and (4.9) hold. Clearly, (4.4) follows from (4.22) and (4.23b). Letting $A_{ii}=b$ for $i=1,\cdots,n$ and $A_{ij}=\bar{v}$ for $i\neq j$, where $\bar{v}$ is such that $f(\bar{v})=b$, gives (4.5). Since $f(b)<b/(n-1)<b$ implies that $b<\bar{v}$, (4.7) holds.

To show (4.6) holds we argue as follows. Let u be a solution of $M(u)=0$ and set

$$s=f(u_1)+\cdots+f(u_n),$$

where $u=(u_1,\cdots,u_n)$. Then from (4.22), for each $i=1,\cdots,n$,

$$s=b-u_i+f(u_i)$$

and so each component of u satisfies the equation

$$f(v)-v+c=0, \tag{4.25}$$

where $c=b-s$. Since (4.23b) implies that $0<f'(v)<1$ for $v>0$, we have that (4.25) has at most one positive solution for any constant c. This then implies that all the components of u must be equal. Let $u_1=\cdots=u_n=\alpha$. Then α satisfies

$$b-\alpha-(n-1)f(\alpha)=0, \tag{4.26}$$

which clearly has exactly one positive solution (see Fig. 4).

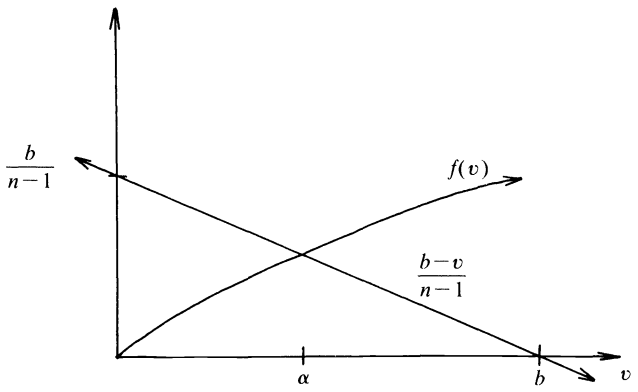


FIG. 4

So the unique feasible equilibrium is $u^*=(\alpha,\cdots,\alpha)$, where α satisfies (4.26).

It remains to show (4.9). Let

$$V_i(\tau)=k_i(\tau\hat{u}^*)=b-(n-1)f(\tau\alpha), \qquad 0\le\tau\le 1,$$

and let

$$h(\tau)=k_i(\hat{V}(\tau))-\tau u_i^*=b-(n-1)f[b-(n-1)f(\tau\alpha)]-\tau\alpha.$$

Then

$$h(0)=b-(n-1)f(b)>0 \tag{4.27}$$

by (4.23d) and

$$\begin{aligned} h(1)&=b-(n-1)f[b-(n-1)f(\alpha)]-\alpha \\ &=b-(n-1)f(\alpha)-\alpha=0 \end{aligned} \tag{4.28}$$

by (4.26). Finally,

$$\begin{aligned} h'(\tau)&=\alpha\left[(n-1)^2 f'(b-(n-1)f(\tau\alpha))\cdot f'(\tau\alpha)-1\right] \\ &<0 \end{aligned} \tag{4.29}$$

for $0\le\tau<1$ by (4.23b) and (4.27)–(4.29) then imply

$$h(\tau)>0 \quad \text{for } 0\le\tau<1,$$

which is equivalent to (4.9). Theorem 4.3 now gives the result. □

We give an example which shows that one must assume condition (4.9) holds for all τ in $[0,1)$. (Recall in the two-species case that one only needs to assume (4.9) at $\tau=0$.)

Example 4.6. Let $M_1,\cdots,M_n$ be given by (4.22) but assume instead that f satisfies

(4.30) (a) $f(0)=0$,
(b) $0<f'(v)<1$ for $v>0$,
(c) $f(v)=b$ has a solution,
(d) $f(b)<b/(n-1)$,
(e) $f'(\alpha)>1/(n-1)$, where α is the solution to $b-v-(n-1)f(v)=0$.

An example of a function satisfying (4.30) is

$$f(v)=\begin{cases}\varepsilon v, & 0\le v\le\dfrac{b}{n-1},\\ \beta\left(v-\dfrac{b}{n-1}\right)+\varepsilon\dfrac{b}{n-1}, & v>\dfrac{b}{n-1},\end{cases}$$

where $1/(n-1)<\varepsilon<1$ and $0<\beta<(1-\varepsilon)/(n-2)$, and then smoothing out the rough edge to make f C^1 (see Fig. 5).

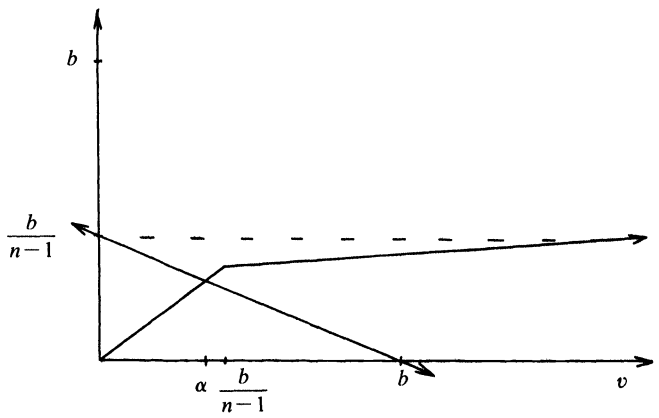


FIG. 5

Conditions (4.4)–(4.7) follow as before, and letting $h(\tau)$ be defined as above, we see that $h(0)>0$ by (4.30d) or that (4.9) holds with $\tau=0$. We show that $h'(1)>0$. Indeed, from what was done above (see (4.29)),

$$h'(1)=\alpha\left[(n-1)^2(f'(\alpha))^2-1\right]>0$$

by (4.30e). Since $h(1)=0$, this implies that $h(\tau)<0$ for $\tau<1$ and close, thus showing that (4.9) doesn't hold for all τ in $[0,1)$.

Finally, we note that a result analogous to Theorem 4.5 holds in the case that the function $f(v)=av^2$, $a>0$, provided $ab<3/4(n-1)$. The proof follows along lines similar to that for Theorem 4.5.

REFERENCES

[1] F. ALBRECHT, H. GATZKE, A. HADDAD AND N. WAX, *The dynamics of two interacting populations*, J. Math. Anal. Appl., 46 (1974), pp. 658–670.

[2] P. N. BROWN, *Decay to uniform states in ecological interactions*, SIAM J. Appl. Math., 38 (1980), pp. 22–37.

[3] K. N. CHEUH, C. C. CONLEY AND J. A. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[4] E. D. CONWAY AND J. A. SMOLLER, *A comparison technique for systems of reaction-diffusion equations*, Comm. Partial Differential Equations, (7) 2 (1977), pp. 679–697.

[5] P. DE MOTTONI AND F. ROTHE, *Convergence to homogeneous equilibrium state for generalized Volterra-Lotka systems with diffusion*, SIAM J. Appl. Math., 37 (1979), pp. 648–663.

[6] B. S. GOH, *Global stability in two species interactions.*, J. Math. Biol., 3 (1976), pp. 313–318.

[7] A. HASTINGS, *Global stability in Lotka-Volterra systems with diffusion*, J. Math. Biol., 6 (1978), pp. 163–168.

[8] _____, *Global stability of two species systems*, J. Math. Biol., 5 (1978), pp. 399–403.

[9] A. RESCIGNO AND I. W. RICHARDSON, *The struggle for life*: I. *Two species*, Bull. Math. Biophys., 29 (1967), pp. 377–388.

[10] J. MAYNARD SMITH, *Models in Ecology*, Cambridge University Press, Cambridge, 1974.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0003 $01.25/0

COMPLETENESS OF DERIVATIVES OF SQUARED SCHRÖDINGER EIGENFUNCTIONS AND EXPLICIT SOLUTIONS OF THE LINEARIZED KdV EQUATION*

ROBERT L. SACHS†

Abstract. Explicit solutions to the Cauchy problem for the linearized KdV equation are constructed when the initial data is integrable. The method is analogous to the Fourier decomposition for a constant coefficient equation and uses the connection between the one-dimensional Schrödinger equation and the KdV equation, as discovered by Gardner, Greene, Kruskal and Miura [2]. An expansion theorem expressing any integrable function in terms of derivatives of squared Schrödinger (generalized) eigenfunctions is proved. These functions evolve according to the linearized KdV equation, hence the expansion of the initial data leads to a generalized solution of the linearized KdV equation. Under suitable restrictions on the initial data, the solution constructed is classical. The proof of the expansion theorem may be interpreted as the skew-adjoint analogue of the more familiar process of simultaneously diagonalizing two self-adjoint operators.

AMS-MOS subject classification (1980). Primary 35C15, 35Q20, 34B25

Key words. linearized KdV equation, Schrödinger eigenfunctions, skew-symmetric operators

1. Introduction. In this paper, we present an explicit solution to the Cauchy problem for the linearized KdV equation:

$$(*) \qquad u_t+u_{xxx}-6(qu)_x=0, \qquad u(x,0)=\phi(x),$$

where $q(x,t)$ is a solution of the KdV equation ((1.5) below). Our method expresses the solution as a superposition of particular solutions and utilizes a completeness theorem which we discuss below. The particular solutions we choose may be thought of as derivatives of $q(x,t)$ with respect to the scattering data for the Schrödinger equation with potential $q(x,t)$. Hence we sketch briefly the inverse scattering method of solving the KdV equation, as discovered by Gardner, Greene, Kruskal and Miura [2].

If we consider the one-dimensional Schrödinger equation with potential $Q(x)$,

$$(1.1) \qquad -\frac{d^2}{dx^2}f+Q(x)f=k^2f$$

and define the Jost solutions $f_\pm(x,k)$ by their asymptotic behavior

$$(1.2) \qquad f_+\sim e^{ikx} \quad \text{as } x\to+\infty, \qquad f_-\sim e^{-ikx} \quad \text{as } x\to-\infty$$

then the relation

$$(1.3) \qquad T(k)f_-(x,k)=f_+(x,-k)+R(k)f_+(x,k),$$

which defines the transmission coefficient $T(k)$ and the reflection coefficient $R(k)$ implies $T(k)$ is meromorphic in $\operatorname{Im}k>0$ with finitely many poles, all on the imaginary axis. The completeness theorem mentioned above expresses any integrable function ϕ in terms of $(f_+^2)'(x,k)$, $(f_-^2)'(x,k)$ and a sum of discrete terms related to the poles of $T(k)$. (While we could use (1.3) to eliminate $(f_-^2)'(x,k)$, it is more convenient not to do

*Received by the editors December 18, 1981, and in revised form May 5, 1982. This research was sponsored by the U.S.A. under contract no. DAAG29-80-C-0041 and by an American Mathematical Society Postdoctoral Research Fellowship.

† Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

so.) We prove the theorem by solving the equation

$$\psi'''-4Q\psi'-2Q'\psi+4k^2\psi'=\phi$$

for ψ' and integrating the "resolvent".

If we now consider a one-parameter family of Schrödinger operators

$$L(t)\equiv-\frac{d^2}{dx^2}+q(x,t), \tag{1.4}$$

where the time evolution of $q(x,t)$ is given by the KdV equation

$$q_t+q_{xxx}-6qq_x=0, \tag{1.5}$$

then it turns out [2] that for any t, $L(t)$ is unitarily equivalent to $L(0)$. This implies that the spectrum of $L(t)$ is invariant. Moreover, the scattering data associated with the operators $L(t)$ evolve in a very simple manner. Zakharov and Faddeev [16] interpret the above facts in the context of completely integrable Hamiltonian systems and show that the eigenvalues of L and k times the logarithm of $|T(k)|^2$ for real k form action variables with appropriate conjugate angle variables. In [16], a formal calculation appears which expresses the infinitesimal variation of $q(x,0)$ in terms of variations of the scattering data. This formula suggests consideration of x-derivatives of the squared eigenfunctions of (1.1) with their induced time dependence as solutions to the linearized KdV equation. The fact that these derivatives satisfy the linearized KdV equation for smooth potentials already appears implicitly in [2, Thm. 3.6]. Reference [12], besides providing an excellent overview of inverse scattering methods for solving evolution equations, presents several results closely related to ours. In particular, Newell derives the orthogonality relations implied by Theorem 2.1 below. However, completeness of these functions is not discussed in [12] or elsewhere, to the best of our knowledge. A related expansion for the Zakharov–Shabat eigenvalue problem appears in Kaup [4]. Discussions of perturbations using the inverse scattering formalism appear in [5], [9] (for the sine-Gordon equation), and [11]; for an application of this result to the problem of water waves in a canal, see [13]. Squared eigenfunctions and their derivatives also play an important role in the theory of the periodic KdV equation [8].

The completeness theorem is proved in §2 below, while in §3, the time evolution of the eigenfunctions and the solution of (*) are discussed. Some of these results appear in the author's doctoral dissertation (New York University, October 1980). The advice and encouragement of his advisor, Jürgen Moser, is gratefully acknowledged.

We also remark that semi-group methods [3] will yield a solution of (*) for L^2 initial data, so for a large class of initial data, we have constructed the "evolution operator" explicitly.

2. L^1-completeness of derivatives of squared Schrödinger eigenfunctions. After introducing some notation and results from the scattering theory of the one-dimensional Schrödinger equation, we state and prove an expansion theorem for derivatives of squared Schrödinger eigenfunctions. (We shall use the term eigenfunction to include generalized eigenfunctions as well as bona fide L^2 solutions.)

Consider the Schrödinger equation

$$-f''(x,k)+Q(x)f(x,k)=k^2f(x,k) \tag{2.1}$$

for k real. Our notation shall be:

$$f'(x,k)=\frac{\partial}{\partial x}f(x,k), \qquad \dot{f}(x,k)=\frac{\partial}{\partial k}f(x,k).$$

We assume the potential $Q(x)$ satisfies

$$\|Q\|_{L_2^1} \equiv \int_{-\infty}^{\infty} (1+x^2)|Q(x)|\,dx < \infty. \tag{2.2}$$

The fundamental discovery of Gardner, Greene, Kruskal and Miura [2], later formulated abstractly by Lax [7], is that if $q(x,t)$ evolves according to the KdV equation, the spectrum of the Schrödinger equation (2.1) with potential $q(x,t)$ remains fixed in t and the associated scattering data evolves in a very simple manner. We shall use this information below, but first introduce some notation and basic facts about scattering theory for (2.1). This information (and much more) may be found in [1].

Let $f_{\pm}(x,k)$ denote the Jost solutions of (2.1), i.e., $f_+(x,k) \sim e^{ikx}$ as $x \to +\infty$, $f_-(x,k) \sim e^{-ikx}$ as $x \to -\infty$, and both satisfy (2.1). The transmission coefficient, $T(k)$, as defined in (1.3) above, is represented in terms of the Wronskian of f_+, f_- by

$$\frac{1}{T(k)} = \frac{1}{2ik}[f_+(x,k), f_-(x,k)] = \frac{f_+' f_- - f_-' f_+}{2ik}. \tag{2.3}$$

Formula (2.3) and the normalization of f_+, f_- imply that $T(k)$ is meromorphic in the upper half-plane $\operatorname{Im} k > 0$ with poles at $k = i\beta_j$, $j = 1, \cdots, N$ where each energy $-\beta_j^2$ is a bound state energy in (2.1). N is finite by a classical estimate assuming $(1+|x|)|Q(x)|$ is integrable. $T(k)$ is also continuous and nonzero for real $k \neq 0$. For notational ease, we also introduce for $j = 1, \cdots, N$ the following pair of functions:

$$F_j(x) = f_+^2(x, i\beta_j), \qquad G_j(x) = c_j f_+(x, i\beta_j) \cdot g_j(x), \tag{2.4}$$

$$\text{where } g_j(x) \equiv \frac{1}{i} \frac{d}{dk}\left[f_-(x,k) - \frac{f_-(x,i\beta_j)}{f_+(x,i\beta_j)} f_+(x,k) \right]_{k=i\beta_j}$$

and c_j is chosen so that $\int_{-\infty}^{\infty} F_j'(x) G_j(x)\,dx = 1$ for $j = 1, \cdots, N$. The expansion theorem mentioned above is:

THEOREM 2.1. *Suppose $Q(x)$ satisfies (2.2). If $\phi(x)$ is continuous and in L^1, then*

$$\phi(x) = \lim_{\varepsilon \downarrow 0} \int_{-\infty + i\varepsilon}^{\infty + i\varepsilon} \frac{dk}{2\pi i k} T^2(k) \cdot \int_{-\infty}^{\infty} K(x,y,k)\phi(y)\,dy \tag{2.5a}$$

$$+ \sum_{j=1}^{N} \int_{-\infty}^{\infty} \left[F_j'(x) G_j(y) - G_j'(x) F_j(y) \right] \phi(y)\,dy$$

and

(2.5b)

$$\phi(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{T^2(k)}{4\pi i k} \left[(f_+^2)'(x,k) f_-^2(y,k) - (f_-^2)'(x,k) f_+^2(y,k) \right] \phi(y)\,dy\,dk$$

$$+ \sum_{j=1}^{N} \int_{-\infty}^{\infty} \left[F_j'(x) G_j(y) - G_j'(x) F_j(y) \right] \phi(y)\,dy,$$

where the kernel $K(x,y,k)$ is defined as

$$K(x,y,k)\equiv\begin{cases}\dfrac{\partial}{\partial x}\left(f_+^2(x,k)f_-^2(y,k)-h(x,k)h(y,k)\right) & \text{for } y\le x,\\ \dfrac{\partial}{\partial x}\left(h(x,k)h(y,k)-f_-^2(x,k)f_+^2(y,k)\right) & \text{for } y\ge x,\end{cases} \tag{2.6}$$

with $h(x,k)\equiv f_+(x,k)f_-(x,k)$.

We remark that for $q=0$, $f_\pm(x,k)=e^{\pm ikx}$ and the above expansion reduces to the ordinary Fourier transform. The latter representation (2.5b) is more convenient in applications while (2.5a) is central to the proof of the theorem. Before presenting the proof, we discuss the choice of the particular kernel $K(x,y,k)$ of (2.6).

If f and g are both C^3 solutions of the Schrödinger equation (2.1) for the same energy k^2, then their product $f\cdot g$ is a solution of the third-order equation

$$\psi'''-4Q\psi'-2Q'\psi=-4k^2\psi'. \tag{2.7}$$

Two linearly independent solutions of (2.1) generate three independent solutions of (2.7), e.g., f^2, fg, g^2. We choose $f_+(x,k)$, $T(k)\cdot f_-(x,k)$. Then, solving the inhomogeneous form of (2.7) for a function $\phi(x)$ by variation of parameters leads to an expression for ψ in terms of ϕ. The kernel $K(x,y,k)$ is the "Green's function" for this problem. Differentiating, we obtain formally

$$\psi'=\left(D^2-4(Q-k^2)-2Q'D^{-1}\right)^{-1}\phi, \tag{2.8}$$

which we integrate as though it were a bona fide resolvent and obtain a multiple of the identity. The calculation of this integral is the content of the following lemma.

LEMMA 2.2. *Let Γ_R be the semicircle in the upper half-plane of radius R traversed from $-R$ to R. Then:*

$$\lim_{R\to\infty}\frac{1}{2\pi i}\int_{\Gamma_R}\frac{T^2(k)}{k}\left\{\int_{-\infty}^{\infty}K(x,y,k)\phi(y)\,dy\right\}dk=\phi(x) \tag{2.9}$$

for all ϕ which are continuous with $\phi\in L^1$.

Postponing the proof of Lemma 2.2 momentarily, we show that Lemma 2.2 implies Theorem 2.1.

Proof of Theorem 2.1 (given Lemma 2.2). Apply Cauchy's theorem. For $\operatorname{Im}k>0$, the integrand in k has poles only at the poles of $T^2(k)$; an easy calculation shows that for $k=i\beta_1,\cdots,i\beta_N$ (recall, the bound state energies are $-\beta_N^2<-\beta_{N-1}^2<\cdots<-\beta_1^2<0$), the pole of $T(k)$ is simple [1]. Thus the integrand we consider has double poles at $k=i\beta_j$, $j=1,\cdots,N$.

For $k=i\beta_j$, there is a constant α_j such that $f_-(x,i\beta_j)=\alpha_j f_+(x,i\beta_j)$. Thus the quantities

(2.10)

$$f_+^2(x,k)f_-^2(y,k)-h(x,k)h(y,k) \quad\text{and}\quad f_+(x,k)f_-(y,k)-f_-(x,k)f_+(y,k)$$

vanish identically in x,y when $k=i\beta_j$.

Let A_j be the residue of $T(k)$ at $k=i\beta_j$. Then the residue of the left-hand side of (2.9) at $k=i\beta_j$ is precisely

$$(2.11)\quad \frac{1}{2\pi i}\frac{A_j^2}{i\beta_j}\frac{d}{dk}\left\{\int_{-\infty}^{\infty}K(x,y,k)\phi(y)\,dy\right\}_{k=i\beta_j}$$

$$=\frac{1}{2\pi i}\frac{A_j^2}{\beta_j}\int_{-\infty}^{\infty}\frac{\partial}{\partial x}\big[f_+^2(x,i\beta_j)\cdot f_+(y,i\beta_j)g_j(y)$$

$$-f_+(x,i\beta_j)g_j(x)f_+^2(y,i\beta_j)_+\big]\phi(y)\,dy,$$

where we recall

$$g_j(x)\equiv\frac{1}{i}\frac{d}{dk}\big[f_-(x,k)-\alpha_j f_+(x,k)\big]_{k=i\beta_j}.$$

The right-hand side of (2.11) comes from replacing $f_-(x,i\beta_j)$ by $\alpha_j f_+(x,i\beta_j)$ after using the remark below (2.10). Now we deform the semicircle to the real k-axis. By our definitions above, the deformation contributes the terms $\Sigma_{j=1}^N\int_{-\infty}^{\infty}[F_j'(x)G_j(y)-G_j'(x)F_j(y)]\phi(y)\,dy$ which appear in the conclusion of Theorem 2.1, in addition to the integration along the real k-axis. We also remark that despite appearances, *neither real k integral has a singularity at $k=0$*. This follows from the fact that either (i) $f_+(x,0)$ and $f_-(x,0)$ are linearly dependent, which by the same remark as above implies that $K(x,y,k)$ vanishes at least linearly in k as $k\to 0$ in $\operatorname{Im}k\geq 0$; or (ii) $f_+(x,0)$ and $f_-(x,0)$ are linearly independent, so by (2.3), $T(k)=\alpha k+o(k)$ as $k\to 0$. In either case, there is no pole at $k=0$. (See [1] for a further discussion of phenomena at $k=0$ in scattering theory.) Thus we have proved (2.5a) assuming Lemma 2.2. To obtain (2.5b), we remark that the difference between the two k-integrals integrates to 0.

The difference between (2.5a)–(2.5b) is precisely

$$\int_{-\infty}^{\infty}\frac{dk}{4\pi ik}\int_{-\infty}^{\infty}\hat{K}(x,y,k)\phi(y)\,dy,$$

where $\hat{K}(x,y,k)=\operatorname{sgn}(x-y)\partial/\partial x[T(k)f_+(x,k)f_-(y,k)-T(k)f_-(x,k)f_+(y,k)]^2$. Using (1.3) to eliminate $f_-(x,k),f_-(y,k)$, we see that $\hat{K}(x,y,k)$ is an even function of k, so the integral vanishes. This proves (2.5b).

To complete the argument, we now prove Lemma 2.2. Consider

$$(2.12)\qquad I_R\equiv\frac{1}{2\pi i}\int_{\Gamma_R}\frac{T^2(k)}{k}\left\{\int_{-\infty}^{\infty}K(x,y,k)\phi(y)\,dy\right\}dk.$$

Write $f_\pm(x,k)=m_\pm(x,k)e^{\pm ikx}$. We shall make use of the following estimates, taken from Deift–Trubowitz [1], which hold for all k in $\operatorname{Im}k\geq 0$

$$(2.13)\quad \text{(i)}\quad |m_\pm(x,k)-1|\leq\exp\left\{\frac{C_1}{|k|}\right\}\frac{C_2}{|k|},$$

$$\text{(ii)}\quad \left|m_+'(x,k)+\int_x^{\infty}e^{2ik(y-x)}Q(y)\,dy\right|\leq C_3/(1+|k|)\quad\text{and similarly for } m_-,$$

$$\text{(iii)}\quad T(k)=1+O\left(\frac{1}{|k|}\right)\quad\text{as } |k|\to\infty.$$

In [1] it is also shown that $m_{\pm}-1$ are Hardy functions; in particular, they are analytic in $\operatorname{Im} k>0$. Recalling the definition of $K(x,y,k)$ given by (2.6), we define:

(2.14)

$$I_R^{(1)} \equiv \frac{1}{2\pi i}\int_{\Gamma_R} dk\,\frac{T^2(k)}{k}\left\{\int_{-\infty}^{x} 2ikm_+^2(x,k)m_-^2(y,k)\,e^{2ik(x-y)}\phi(y)\,dy \right.$$
$$\left. +\int_x^{\infty} 2ikm_-^2(x,k)m_+^2(y,k)e^{-2ik(x-y)}\phi(y)\,dy\right\},$$

$$I_R^{(2)} \equiv \frac{1}{2\pi i}\int_{\Gamma_R} dk\,\frac{T^2(k)}{k}\left\{\int_{-\infty}^{x}\left[2m_+(k,x)m'_+(x,k)m_-^2(y,k)e^{2ik(x-y)}\right.\right.$$
$$\left. -h'(x,k)h(y,k)\right]\phi(y)\,dy$$
$$+\int_x^{\infty}\left[h'(x,k)h(y,k)\right.$$
$$\left.\left. -2m_-(x,k)m'_-(x,k)m_+^2(y,k)\cdot e^{-2ik(x-y)}\right]\phi(y)\,dy\right\}.$$

Thus $I_R = I_R^{(1)} + I_R^{(2)}$.

By estimates (i), (ii), (iii) in (2.13), since $I_R^{(2)}$ contains terms which have a factor $m'_{\pm}(x,k)$ and $m_{\pm}(x,k)$ is uniformly bounded for $|k|>c>0$, we have the estimate

$$|I_R^{(2)}| \le C\,\frac{\|\phi\|_{L^1}}{R} \quad \text{for all } R \ge R_0 \text{ sufficiently large,} \tag{2.15}$$

where C is independent of R.

Morover, by (2.13) (i), (iii), $m_{\pm}(x,k)=1+O(1/R)$. $T(k)=1+O(1/R)$ for $|k|=R$ so

$$I_R^{(1)} = \frac{1}{2\pi i}\int_{\Gamma_R}\frac{dk}{k}\left\{\int_{-\infty}^{x} 2ik\,e^{2ik(x-y)}\phi(y)\,dy \right.$$
$$\left. +\int_x^{\infty} 2ik\,e^{-2ik(x-y)}\phi(y)\,dy\right\} + O\left(\frac{1}{R}\right).$$

The first terms converge to $\phi(x)$ as in the usual proof of Fourier completeness [15] and the lemma is proved by taking $R\to\infty$.

Remarks. (i) The expansion theorem above bears a strong resemblance to that of the Fourier transform for L^1 functions. However, since the underlying process is the "simultaneous diagonalization" of the two skew operators d/dx and $-(d/dx)^3+2(d/dx)Q+2Qd/dx$, the analogue of the Fourier L^2 theory is not obvious if $Q\not\equiv 0$. If we define

$$\hat{\phi}_{\pm}(k) = \int_{-\infty}^{\infty}\phi(y)f_{\pm}^2(y,k)\,dy, \tag{2.16}$$

the natural version of the Plancheral formula in this case relates the skew bilinear form $\int_{-\infty}^{\infty}\psi'(x)\phi(x)$ to the standard symplectic pairing

$$\begin{pmatrix}\hat{\psi}_+(k)\\ \hat{\psi}_-(k)\end{pmatrix}, \qquad \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}\hat{\phi}_+(k)\\ \hat{\phi}_-(k)\end{pmatrix}.$$

(ii) We also note that the skew operator $-D^3+2DQ+2QD$ ($D=d/dx$) was used by Lenard to recursively generate the KdV conservation laws [2]. It seems to play a crucial role in many aspects of the KdV theory—e.g., it is useful in proving that the integrals are in involution.

3. Application of Theorem 2.1 to the Cauchy problem for the linearized KdV equation. The expansion of $\phi(x)$ given by Theorem 2.1 will lead directly to a method for solving the Cauchy problem for the linearized KdV equation:

$$(*) \qquad u_t+u_{xxx}-6(qu)_x=0, \qquad u(x,0)=\phi(x)$$

with ϕ satisfying the hypotheses of Theorem 2.1. As the potential $q(x,t)$ in the Schrödinger equation

$$-f''+qf=k^2f$$

evolves according to the KdV equation, the corresponding eigenfunctions evolve in time. Gardner, Greene, Kruskal and Miura [2] observed that the squares of the eigenfunctions satisfy the formal adjoint of the linearized KdV equation (which they called "the associated linear equation"), namely

$$(3.1) \qquad v_t+v_{xxx}-6qv_x=0$$

from which it follows that $u\equiv v_x$ satisfies

$$(3.2) \qquad u_t+u_{xxx}-6(qu)_x=0 \quad \text{(the linearized KdV equation)}.$$

In view of this fact, the expansion of Theorem 2.1 may be extended to include the time evolution of the eigenfunctions. As we shall see below, this extension is the solution of the Cauchy problem (*).

We begin by developing some necessary preliminary facts.

LEMMA 3.1. *The functions*

$$g_j(x)\equiv\frac{1}{i}\frac{d}{dk}\left[f_-(x,k)-\alpha_j f_+(x,k)\right], \qquad k=i\beta_j$$

are (unbounded) solutions of the Schrödinger equation (2.1) *with* $k^2=-\beta_j^2$.

Proof. Differentiating (2.1) with respect to k, we obtain an equation for $(d/dk)f_\pm(x,k)=\dot{f}_\pm(x,k)$:

$$(3.3) \qquad \dot{f}_\pm''=(q-k^2)\dot{f}_\pm-2kf_\pm.$$

Consider $g_j=[\dot{f}_--\alpha_j\dot{f}_+]/i$ at $k=i\beta_j$: $g_j(x)$ satisfies

$$\begin{aligned} g_j'' &= \left(q+\beta_j^2\right)g_j-2\beta_j\left(f_-(x,i\beta_j)-\alpha_j f_+(x,i\beta_j)\right) \\ &= \left(q+\beta_j^2\right)g_j \quad \text{by our choice of } \alpha_j. \end{aligned}$$

Remark. $g_j(k)$ is exponentially increasing as $|x|\to\infty$; however, the product $f_+(x,i\beta_j)g_j(x)$ is bounded.

We now discuss the result in [2] mentioned above, and sketch the proof.

LEMMA 3.2. (*cf.* [2, *eq.* (2.19)]). *Let ψ be a solution of the Schrödinger equation* (2.1) *with potential $q(x,t)$ evolving according to the KdV equation. Then the function*

$$(3.4) \qquad R\equiv\psi_t+\psi_{xxx}-3(q+k^2)\psi_x$$

is also a solution of the Schrödinger equation with potential $q(x,t)$.

Sketch of the proof. Use the equation $\psi''=(q-k^2)\psi$ to express q in terms of ψ. Substitute into the KdV equation and simplify the resulting expression, obtaining the equation $R''-(q-k^2)R=0$ after eliminating a factor of ψ.

Remark. The chief use of Lemma 3.2 is to show that R, for suitable eigenfunctions ψ, is in fact 0. In particular, we have:

COROLLARY 3.3. *The expression R vanishes if we choose any of the following eigenfunctions for ψ:*

(i) $\tilde{f}_{\pm}(x,k,t)\sim\exp\{\pm i(kx+4k^3t)\}$ *as* $x\to\pm\infty$, *t fixed,*

(ii) $\tilde{f}_{+}(x,i\beta_j,t)$,

(iii) $g_j(x,t)\equiv\frac{1}{i}\frac{d}{dk}\left[\tilde{f}_{-}(x,k,t)-\alpha_j\tilde{f}_{+}(x,k,t)\right]_{k=i\beta_j}$.

Sketch of proof (see [2, Thm. 3.6]). Consider R in each case. By our choice of asymptotics, we have $R\equiv 0$ since no other solution of the Schrödinger equation with that type of decay exists, namely $R\to 0$ as $x\to\pm\infty$ in (i), $\exp\{|\beta_j x-4\beta_j^3 t|\}R\to 0$ in (ii), (iii) as $|\beta_j x-4\beta_j^3 t|\to\infty$. Thus from the spectral theory of the Schrödinger equation, $R\equiv 0$ in each case.

LEMMA 3.4. *Suppose ψ_1,ψ_2 are to (not necessarily independent) solutions of the Schrödinger equation for the same eigenvalue. If $\psi_t+\psi_{xxx}-3(q+k^2)\psi_x\equiv 0$ for $\psi=\psi_j$, $j=1,2$, then the product $\psi_1\psi_2$ is a solution of the adjoint equation:*

$$v_t+v_{xxx}-6qv_x=0.$$

Proof. A direct calculation:

$$\begin{aligned}(\psi_1\psi_2)_t&+(\psi_1\psi_2)_{xxx}-6q(\psi_1\psi_2)_x\\&=\psi_1(\psi_{2_t}+\psi_{2_{xxx}}-6q\psi_{2_x})+\psi_2(\psi_{1_t}+\psi_{1_{xxx}}-6q\psi_{1_x})+3\psi_{1_x}\psi_{2_{xx}}+3\psi_{1_{xx}}\psi_{2_x}\\&=\psi_1(\psi_{2_t}+\psi_{1_{xxx}}-3(q+k^2)\psi_{2_x})+\psi_2(\psi_{1_t}+\psi_{1_{xxx}}-3(q+k^2)\psi_{1_x})=0,\end{aligned}$$

where we have used the Schrödinger equation to eliminate the second derivatives.

Remarks. (i) An alternate derivation of these facts may be given using the following idea of Tanaka [14]: The KdV equation may be written in the Lax [7] form

$$\frac{dL}{dt}=[B,L], \tag{3.5}$$

where $L(t)$ is the operator $-d^2/dx^2+q(x,t)$ and $B(t)$ is the skew operator

$$-4\frac{d^3}{dx^3}+3q(x,t)\frac{d}{dx}+3\frac{d}{dx}q(x,t).$$

The time derivative of the Schrödinger equation (2.1) together with (3.5) implies

$$L(f_t-Bf)=k^2(f_t-Bf). \tag{3.6}$$

Choosing $f=f_+(x,k,t)\sim e^{ikx}$ as $x\to+\infty$ for t fixed and analyzing the asymptotic behavior of f_t-Bf as $x\to+\infty$ for t fixed implies

$$(f_+)_t-B(f_+)=4(ik)^3(f_+) \tag{3.7}$$

so $\psi(x,k,t)\equiv e^{4ik^3t}\cdot f_+(x,k,t)$ satisfies

$$\psi_t-B\psi=0,\qquad \psi\sim e^{ikx+4k^3t}\quad\text{as } x\to+\infty\quad\text{for } t \text{ fixed} \tag{3.8}$$

(a similar argument holds for $e^{-4ik^3t}\cdot f_-(x,k,t)$). Using the Schrödinger equation (2.1), a simple calculation, as in Lemma 3.4 above, shows that products (with the same values of k^2) of solutions of (3.8) satisfy the adjoint equation

$$v_t+v_{xxx}-6qv_x=0.$$

(ii) If $q(x,t)$ is a classical solution of the KdV equation, the formal calculations above are sensible—i.e., the eigenfunctions possess the necessary derivatives. This follows from the inhomogeneous form of the Schrödinger equation which these derivatives satisfy.

Using these facts, we make the following definitions (extending (2.4) to include time dependence):

$$
\begin{aligned}
&\tilde{f}_+(x,k,t)\sim e^{ikx+4ik^3t} \quad \text{as } x\to+\infty, \quad t \text{ fixed},\\
&\tilde{f}_-(x,k,t)\sim e^{-ikx-4ik^3t} \quad \text{as } x\to-\infty, \quad t \text{ fixed},\\
&\tilde{f}_\pm \text{ satisfy (2.1) with potential } q(x,t),\\
&\tilde{F}_j(x,t)=f_+^2(x,i\beta_j,t),\\
&\tilde{G}_j(x,t)=c_j\tilde{f}_+(x,i\beta_j,t)\cdot\tilde{g}_j(x,t).
\end{aligned}
\tag{3.9}
$$

The obvious candidate for the solution of (*) is the function $u(x,t)$ defined as follows:

$$
\begin{aligned}
u(x,t)\equiv&\int_{-\infty}^{\infty}\frac{dkT^2(k)}{4\pi ik}\left[(\tilde{f}_+^2)'(x,k,t)\hat{\phi}_-(k)-(\tilde{f}_-^2)'(x,k,t)\hat{\phi}_+(k)\right]\\
&+\sum_{j=1}^{N}\int_{-\infty}^{\infty}\left[F_j'(x,t)G_j(y,0)-G_j'(x,t)F_j(y,0)\right]\phi(y)\,dy,
\end{aligned}
\tag{3.10}
$$

where $\hat{\phi}_\pm(k)$ are defined in (2.16) above. By Lemma 3.4, all of the functions of (x,t) appearing on the right-hand side of (3.10) satisfy the linearized KdV equation. Thus we have proved:

LEMMA 3.5. *The function $u(x,t)$ defined by* (3.10) *above satisfies the linearized KdV equation*:

$$
u_t+u_{xxx}-6(qu)_x=0, \qquad u(x,0)=\phi(x) \tag{*}
$$

in the sense of distributions.

To prove that $u(x,t)$ is a classical solution of (*) for $t>0$, we need some additional smoothness and decay on $\phi(x)$, the initial data. The situation is completely analogous with the Fourier transform solution of the Airy equation:

$$
w_t+w_{xxx}=0. \tag{3.11}
$$

The x-decay of ϕ is needed to have $\hat{\phi}_\pm(k)$ be differentiable while smoothness of ϕ relates to integrability of $k^\alpha\hat{\phi}_\pm(k)$ for $0\le\alpha\le2$. As in the case of (3.11), one can prove the following:

THEOREM 3.6. *The function $u(x,t)$, given by* (3.20) *above, is a classical solution of the linearized KdV equation for $t>0$ if*

(i) *$\phi(x)$ has four continuous derivatives,*

(ii) *as $|x|\to\infty$, $\partial_x^r\phi(x)=O(|x|^{-4})$ for $r=0,1,2,3,4$.*

Sketch of proof. Using (i) and the definition of $\hat{\phi}_\pm(k)$, we integrate by parts four times with respect to x, which implies that as $|k|\to\infty$, $\hat{\phi}_\pm(k)=O(|k|^{-4})$. Also, by (ii), $\hat{\phi}_\pm(k)$ are C^2. Thus $u(x,t)$ has two continuous derivatives with respect to x. As in Murray [10], the factor k^2 is written as $(x+12k^2t)/12t-x/12t$, where the first term is a multiple of the k-derivative of the exponentials $e^{\pm 2ik(x+4k^2t)}$. Integrating by parts in k, we find that $u(x,t)$ has four continuous x-derivatives. Repeating the argument, we

obtain six x-derivatives. From the linearized KdV equation, this implies u_t is continuous, hence it is a classical solution. The only difference between this case and the Airy equation (3.11) is the presence of the added factors

$$\tilde{m}_{\pm}(x,k,t)=\tilde{f}_{\pm}(x,k,t)e^{\mp 2ik(x+4k^2t)}$$

and these do not affect the necessary estimates.

A fuller discussion and an analysis of asymptotic behavior is presented for the N-soliton linearization in [13], where the perturbation theory for the problem of water waves in a canal is discussed. The KdV equation was first derived to model precisely this situation [6].

REFERENCES

[1] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.

[2] C. GARDNER, J. GREENE, M. KRUSKAL AND R. MIURA, *Korteweg-de Vries equation and generalizations. VI. methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97–133.

[3] T. KATO, *Linear evolution equations of 'hyperbolic' type*, J. Fac. Sci. Univ. Tokyo, Sec. I, 17 (1970), pp. 241–258.

[4] D. KAUP, *Closure of the squared Zakharov–Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.

[5] J. P. KEENER AND D. W. MCLAUGHLIN, *Solutions under perturbations*, Phys. Rev. A., 16 (1977), pp. 777–790.

[6] D. J. KORTEWEG AND G. DE VRIES, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary wave*, Phil. Mag., 39 (1895), pp. 422–433.

[7] P. D. LAX, *Integrals of non-linear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[8] H. P. MCKEAN AND E. TRUBOWITZ, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29 (1976), pp. 143–226.

[9] D. W. MCLAUGHLIN AND A. C. SCOTT, *Perturbation analysis of fluxon dynamics*, Phys. Rev. A., 18 (1978), pp. 1652–1680.

[10] A. C. MURRAY, *Existence and regularity for solutions of the Korteweg-de Vries equation*, Arch. Rational Mech. Anal., 71 (1979), pp. 143–175.

[11] A. C. NEWELL, *Near integrable systems, nonlinear tunnelling and solitons in slowly changing media*, in Nonlinear Evolution Equations Solvable by the Spectral Transform, F. Calogero, ed., Research Notes in Mathematics, 26, Pitman, San Francisco, 1978.

[12] _____, *The inverse scattering transform*, in Solitons, R. K. Bollough and P. J. Caudrey, eds., Springer-Verlag, New York, 1980.

[13] R. L. SACHS, *A justification of the* KdV *approximation to first order in the case of N-tuple water waves*, MRC Tech. Summary Rep. 2208, Math. Res. Center, Univ. of Wisconsin, Madison, 1981.

[14] S. TANAKA, *Korteweg-de Vries equation: Construction of solutions in terms of scattering data*, Osaka J. Math., 11 (1974), pp. 49–59.

[15] E. C. TITCHMARSH, *An Introduction to the Theory of Fourier Integrals*, 2nd edition, Clarendon Press, Oxford, 1948.

[16] V. E. ZAKHAROV AND L. D. FADDEEV, *Korteweg-de Vries equation: A completely integrable Hamiltonian System*, Functional Anal. i. Prilozhen, 5 (1971), pp. 18–27; Functional Anal. Appl., 5 (1971), pp. 280–287.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0004 $01.25/0

SOLUTIONS TO THE EQUATIONS OF ONE-DIMENSIONAL VISCOELASTICITY IN BV*

JONG UHN KIM†

Abstract. The initial value problem associated with the equations of one-dimensional viscoelasticity of the rate type is studied. The system of equations is linearized and the solutions to the resulting linear system are completely analyzed in the L^1-setting by the method of Fourier transform. This enables us to establish the global existence of solutions to the original nonlinear system with small initial data in $L^1 \cap \mathrm{BV}$. As a by-product, we also obtain precise results on asymptotic behavior of solutions.

Key words. equations of one-dimensional viscoelasticity of the rate type, linear system, Fourier transform, functions of bounded variation, global existence, asymptotic behavior, nonlinear system

Introduction. In this paper we study the equations of one-dimensional viscoelasticity of the rate type,

$$(0.1)^* \qquad \begin{aligned} u_t^* &= v_x, \\ v_t &= \sigma^*(u^*)_x + v_{xx}, \qquad (x,t) \in (-\infty,\infty)\times[0,\infty), \end{aligned}$$

with initial conditions

$$(0.2)^* \qquad \begin{aligned} u^*(x,0) &= u_0^*(x), \\ v(x,0) &= v_0(x), \qquad x \in (-\infty,\infty), \end{aligned}$$

where v is velocity and u^* is the deformation gradient (the inverse of density). We will discuss this Cauchy problem when $(u_0^*(x), v_0(x))$ are close to a stationary state $(U,0)$, where U is a positive constant. Hence we introduce new functions $u \overset{\text{def}}{=} u^* - U$ and $\sigma(\xi) \overset{\text{def}}{=} \sigma^*(\xi + U)$. Then (0.1)* and (0.2)* reduce to

$$(0.1) \qquad \begin{aligned} u_t &= v_x, \\ v_t &= \sigma(u)_x + v_{xx} \end{aligned}$$

and

$$(0.2) \qquad \begin{aligned} u(x,0) &= u_0(x) \overset{\text{def}}{=} u_0^*(x) - U, \\ v(x,0) &= v_0(x). \end{aligned}$$

From physical considerations, $u^* = u + U$ should be positive, i.e., $u > -U$.

The function σ will be assumed to be C^2-smooth and $\dot{\sigma}(0) > 0$, a condition motivated by mechanics. The conservation of mass is expressed by the first equation of (0.1) and the balance of linear momentum by the second one. For the initial-boundary value problem for (0.1), existence theorems have been established in several function classes (see [1], [2] and [5]).

*Received by the editors July 11, 1981 and in final form July 16, 1982. This research has been supported in part by the National Science Foundation under contract ENG-CME80-23824.

†Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. Presently at: Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

For the hyperbolic system of conservation laws

$$u_t=v_x, \qquad v_t=\sigma(u)_x \tag{0.3}$$

associated with (0.1), it is known that the natural function space is the class BV of functions of bounded variation. More precisely, in the case of initial data with small total variation, the existence of global BV solutions to (0.3), (0.2) was established by using Glimm's scheme [4]. From the physical viewpoint, (0.3) can be visualized as the limit of (0.1) as viscosity vanishes. So it is natural to assume that for initial data of small total variation, global solutions of (0.1), (0.2) should exist in the same class of functions. (For more details, see [3].)

The intent of this paper is to verify this conjecture. The main result is Theorem 2.1. As explained in [3], it is very difficult to apply Glimm's scheme to our problem. Therefore, we shall approach the problem differently. First we consider the linearized version of (0.1):

$$u_t=v_x, \qquad v_t=\dot\sigma(0)u_x+v_{xx}; \tag{0.4}$$

and then regard (0.1) as a perturbed equation with an extra term $\sigma(u)_x-\dot\sigma(0)u_x$. This approach is straightforward but effective, since we can study all the important properties of global solutions to (0.4), (0.2) so as to dispense with a priori estimates for the nonlinear problem which are difficult in the L^1-setting.

We shall begin to analyze (0.4) in §1 and return to (0.1) in §2.

1. Linearized equation. Without loss of generality, we assume $\dot\sigma(0)=1$. Applying a Fourier transform with respect to the space variable x, (0.4) yields

$$\frac{\partial}{\partial t}\hat Y(\xi,t)=\hat A(\xi)\hat Y(\xi,t), \tag{1.1}$$

where

$$\hat Y(\xi,t)=\begin{pmatrix}\hat u(\xi,t)\\ \hat v(\xi,t)\end{pmatrix} \quad\text{and}\quad \hat A(\xi)=\begin{pmatrix}0, & i\xi\\ i\xi, & -\xi^2\end{pmatrix}.$$

Throughout this paper we denote by $\|\cdot\|$ the L^1-norm of a function as well as the total variation of a measure. We want to obtain precise estimates of $\|\mathcal{F}^{-1}[e^{t\hat A(\xi)}]_{ij}\|$, $i,j=1,2$, where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. To simplify notation, we shall denote $e^{t\hat A(\xi)}$ and $\mathcal{F}^{-1}[e^{t\hat A(\xi)}]$ by $\hat G$ and G, respectively. It is easy to express each component $\hat G_{ij}$ of $\hat G$ explicitly:

$$\hat G_{11}=\begin{cases}\dfrac{1}{2}\left(1+\dfrac{\xi^2}{\sqrt{\xi^4-4\xi^2}}\right)e^{\lambda_1 t}+\dfrac{1}{2}\left(1-\dfrac{\xi^2}{\sqrt{\xi^4-4\xi^2}}\right)e^{\lambda_2 t} & \text{for } |\xi|>2,\\[2ex] \dfrac{1}{2}\left(1+\dfrac{\xi}{i\sqrt{4-\xi^2}}\right)e^{\tilde\lambda_1 t}+\dfrac{1}{2}\left(1-\dfrac{\xi}{i\sqrt{4-\xi^2}}\right)e^{\tilde\lambda_2 t} & \text{for } |\xi|<2,\end{cases} \tag{1.2}$$

where

$$\lambda_1=\frac{1}{2}\left(-\xi^2+\sqrt{\xi^4-4\xi^2}\right), \qquad \tilde\lambda_1=\frac{1}{2}\left(-\xi^2+i\xi\sqrt{4-\xi^2}\right),$$
$$\lambda_2=\frac{1}{2}\left(-\xi^2-\sqrt{\xi^4-4\xi^2}\right), \qquad \tilde\lambda_2=\frac{1}{2}\left(-\xi^2-i\xi\sqrt{4-\xi^2}\right);$$

$$\hat{G}_{12}=\hat{G}_{21}=\begin{cases}\dfrac{i\xi}{\sqrt{\xi^4-4\xi^2}}\left(e^{\lambda_1 t}-e^{\lambda_2 t}\right) & \text{for } |\xi|>2,\\ \dfrac{1}{\sqrt{4-\xi^2}}\left(e^{\tilde{\lambda}_1 t}-e^{\tilde{\lambda}_2 t}\right) & \text{for } |\xi|<2;\end{cases} \tag{1.3}$$

$$\hat{G}_{22}=\begin{cases}\dfrac{1}{2}\left(1-\dfrac{\xi^2}{\sqrt{\xi^4-4\xi^2}}\right)e^{\lambda_1 t}+\dfrac{1}{2}\left(1+\dfrac{\xi^2}{\sqrt{\xi^4-4\xi^2}}\right)e^{\lambda_2 t} & \text{for } |\xi|>2,\\ \dfrac{1}{2}\left(1-\dfrac{\xi}{i\sqrt{4-\xi^2}}\right)e^{\tilde{\lambda}_1 t}+\dfrac{1}{2}\left(1+\dfrac{\xi}{i\sqrt{4-\xi^2}}\right)e^{\tilde{\lambda}_2 t} & \text{for } |\xi|<2.\end{cases} \tag{1.4}$$

For $|\xi|=2$, $\hat{G}_{ij}(\pm 2,t)=\lim_{s\to\pm 2}\hat{G}_{ij}(s,t)$, $i,j=1,2$.
Let us set $\hat{H}_1(\xi,t)=\hat{G}_{11}-e^{-t}$ and $H_1(x,t)=\mathcal{F}^{-1}\hat{H}_1(\xi,t)$. Then

LEMMA 1.1. *$H_1(x,t)$, $(\partial/\partial x)H_1(x,t)\in C([0,\infty);L^1(R_x))$ and*

$$\|H_1(x,t)\|\le M_1, \tag{1.5}$$

$$\left\|\frac{\partial}{\partial x}H_1(x,t)\right\|\le M_1(t+1)^{-1/2} \tag{1.6}$$

hold for all $t\ge 0$, for some positive constant M_1.

Proof. Set

$$\hat{H}_2(\xi,t)=\hat{H}_1(\xi,t)-e^{-\xi^2 t/2}\cos(\xi t)$$

and

$$H_2(x,t)=\mathcal{F}^{-1}\hat{H}_2.$$

We shall show that $H_2(x,t)$, $(\partial/\partial x)H_2(x,t)\in C([1,\infty);L^1(R_x))$,

$$\|H_2(x,t)\|\le M_2 t^{-1/6} \quad \text{and} \tag{1.7}$$

$$\left\|\frac{\partial}{\partial x}H_2(x,t)\right\|\le M_2 t^{-3/4} \tag{1.8}$$

hold for all $t\ge 1$, for some positive constant M_2. Let η,ρ be positive constants (to be fixed later on) such that $0<\rho\ll 2\ll\eta<\infty$. Then we have the following estimates:

$$\int_{|\xi|\ge\eta}\left|\hat{H}_2(\xi,t)\right|d\xi\le L(te^{-\alpha t}+e^{-\alpha t}) \quad \text{for all } t\ge 1, \tag{1.9}$$

where L and α are positive generic constants which depend only on η and ρ,

$$\int_{|\xi|\le\rho}\left|\hat{H}_2(\xi,t)\right|d\xi\le Lt^{-1} \quad \text{for all } t\ge 1, \tag{1.10}$$

$$\int_{|\xi|\ge\eta}\left|\frac{\partial}{\partial\xi}\hat{H}_2(\xi,t)\right|^2 d\xi\le L(t^2+1)e^{-\alpha t} \quad \text{for all } t\ge 1 \tag{1.11}$$

$$\int_{|\xi|\le\rho}\left|\frac{\partial}{\partial\xi}\hat{H}_2(\xi,t)\right|^2 d\xi\le Lt^{1/2} \quad \text{for all } t\ge 1. \tag{1.12}$$

Since $\hat{H}_2(\xi,t)\in C^\infty([\rho,\eta]\times R)$ and for some $\alpha>0$, $\lambda_i\le-\alpha$ for $2\le|\xi|\le\eta$, $\mathrm{Re}\tilde{\lambda}_i\le-\alpha$ for $\rho\le|\xi|\le 2$, $i=1,2$, we conclude that

$$\|\hat{H}_2(\xi,t)\|\le Lt^{-1} \quad\text{and}\quad \left\|\frac{\partial}{\partial\xi}\hat{H}_2(\xi,t)\right\|_{L^2}\le Lt^{1/4} \quad\text{for all } t\ge 1. \tag{1.13}$$

In the meantime, we have

$$\begin{aligned}\|H_2(x,t)\| &= \int_{|x|\le T}|H_2(x,t)|\,dx+\int_{|x|\ge T}\frac{1}{|x|}|xH(x,t)|\,dx\\ &\le 2T\|\hat{H}_2(\xi,t)\|+\frac{\sqrt{2}}{\sqrt{T}}\left\|\frac{\partial}{\partial\xi}\hat{H}_2(\xi,t)\right\|_{L^2} \quad\text{for all } T>0.\end{aligned} \tag{1.14}$$

Hence, by taking $T=t^{5/6}$, $\|H_2(x,t)\|\le M_2t^{-1/6}$ holds for all $t\ge 1$. By (1.14) and Lebesgue's dominated convergence theorem, it is easy to see that $H_2(x,t)\in C([1,\infty);L^1(R_x))$. Similar estimates can be obtained for $(\partial/\partial x)H_2(x,t)$:

$$\int_{|\xi|\ge\eta}\left|\xi\hat{H}_2(\xi,t)\right|^2d\xi\le L(t^2+1)e^{-\alpha t} \quad\text{for all } t\ge 1, \tag{1.15}$$

$$\int_{|\xi|\le\rho}\left|\xi\hat{H}_2(\xi,t)\right|^2d\xi\le Lt^{-5/2} \quad\text{for all } t\ge 1, \tag{1.16}$$

$$\int_{|\xi|\ge\eta}\left|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_2(\xi,t)\right)\right|^2d\xi\le L(t^2+1)e^{-\alpha t} \quad\text{for all } t\ge 1, \tag{1.17}$$

$$\int_{|\xi|\le\rho}\left|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_2(\xi,t)\right)\right|^2d\xi\le Lt^{-1/2} \quad\text{for all } t\ge 1. \tag{1.18}$$

By (1.15) to (1.18), we have

$$\|\xi\hat{H}_2(\xi,t)\|_{L^2}\le Lt^{-5/4} \quad\text{and}\quad \left\|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_2(\xi,t)\right)\right\|_{L^2}\le Lt^{-1/4} \quad\text{for all } t\ge 1. \tag{1.19}$$

In combination with

$$\begin{aligned}\left\|\frac{\partial}{\partial x}H_2(x,t)\right\| &= \int_{|x|\le T}\left|\frac{\partial}{\partial x}H_2(x,t)\right|dx+\int_{|x|\ge T}\frac{1}{|x|}\left|x\frac{\partial}{\partial x}H_2(x,t)\right|dx\\ &\le\sqrt{2T}\|\xi\hat{H}_2(\xi,t)\|_{L^2}+\frac{\sqrt{2}}{\sqrt{T}}\left\|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_2(\xi,t)\right)\right\|_{L^2} \quad\text{for all } T>0,\end{aligned} \tag{1.20}$$

(1.19) yields

$$\left\|\frac{\partial}{\partial x}H_2(x,t)\right\|\le M_2t^{-3/4} \quad\text{for all } t\ge 1, \tag{1.21}$$

by taking $T=t$. The argument for $\partial/\partial xH_2\in C([1,\infty);L^1(R_x))$ is similar. Hence it follows immediately that $H_1(x,t)$, $(\partial/\partial x)H_1(x,t)\in C([1,\infty);L^1(R_x))$ and for some positive constant M_3, $\|H_1(x,t)\|\le M_3$ and $\|\partial/\partial xH_1(x,t)\|\le M_3t^{-1/2}$ hold for all $t\ge 1$, since

$$\mathcal{F}^{-1}\left[e^{-\xi^2t/2}\cos(\xi t)\right]=\frac{1}{2}\frac{1}{\sqrt{2\pi t}}\left(e^{-(x+t)^2/2t}+e^{-(x-t)^2/2t}\right). \tag{1.22}$$

We can show directly that

$$(1.23)\qquad \|\hat{H}_1(\xi,t)\|\le L,\qquad \left\|\frac{\partial}{\partial\xi}\hat{H}_1(\xi,t)\right\|\le L,$$

$$\|\xi\hat{H}_1(\xi,t)\|_{L^2}\le L,\qquad \left\|\frac{\partial}{\partial\xi}(\xi\hat{H}_1(\xi,t))\right\|_{L^2}\le L \quad \text{for all } t\in[0,2].$$

By a similar procedure, it is easy to see that $H_1(x,t)$, $(\partial/\partial x)H_1(x,t)\in C([0,2];L^1(R_x))$ and so the proof is complete. □

Next we define

$$\hat{H}_3(\xi,t)=\hat{G}_{12}-ie^{-\xi^2t/2}\sin(\xi t),$$

$$\hat{H}_4(\xi,t)=\xi\hat{G}_{12}-ie^{-t}+ie^{-\xi^2t},$$

$$\hat{H}_5(\xi,t)=\xi\hat{G}_{12}-ie^{-t}-i\xi e^{-\xi^2t/2}\sin(\xi t)$$

and

$$H_i(x,t)=\mathscr{F}^{-1}\hat{H}_i(\xi,t),\qquad i=3,4,5.$$

Then we have

LEMMA 1.2. *$G_{12}(x,t)\in C([0,\infty);L^1(R_x))$, $H_4(x,t)\in C([0,2];L^1(R_x))$, $H_5(x,t)\in C([1,\infty);L^1(R_x))$ and $(\partial/\partial x)H_5(x,t)\in C((0,\infty);L^1(R_x))$. Furthermore,*

$$(1.24)\qquad \|G_{12}(x,t)\|\le M_4 \quad \textit{for all } t\ge 0,$$

$$(1.25)\qquad \|H_5(x,t)\|\le M_4t^{-2/3} \quad \textit{for all } t\ge 1 \quad \textit{and}$$

$$(1.26)\qquad \left\|\frac{\partial}{\partial x}H_5(x,t)\right\|\le M_4(t^{5/4}+t^{1/2})^{-1} \quad \textit{for all } t>0,$$

where M_4 is a positive constant.

Proof. We first display the following list of estimates which can be obtained by elementary computations:

$$(1.27)\qquad \int_{|\xi|\ge\eta}|\hat{H}_3|^2\,d\xi\le Le^{-\alpha t},\qquad \int_{|\xi|\le\rho}|\hat{H}_3|^2\,d\xi\le Lt^{-3/2} \quad \text{for all } t\ge 1,$$

$$(1.28)\qquad \int_{|\xi|\ge\eta}\left|\frac{\partial}{\partial\xi}\hat{H}_3\right|^2 d\xi\le L(t^2+1)e^{-\alpha t},\qquad \int_{|\xi|\le\rho}\left|\frac{\partial}{\partial\xi}\hat{H}_3\right|^2 d\xi\le Lt^{1/2} \quad \text{for all } t\ge 1,$$

$$(1.29)\qquad \|\hat{G}_{12}(\xi,t)\|_{L^2}\le L,\qquad \left\|\frac{\partial}{\partial\xi}\hat{G}_{12}(\xi,t)\right\|_{L^2}\le L \quad \text{for all } t\in[0,2],$$

$$(1.30)\qquad \|\hat{H}_4(\xi,t)\|_{L^2}\le L,\qquad \left\|\frac{\partial}{\partial\xi}\hat{H}_4(\xi,t)\right\|_{L^2}\le L \quad \text{for all } t\in[0,2],$$

$$(1.31)\qquad \int_{|\xi|\ge\eta}|\hat{H}_5(\xi,t)|\,d\xi\le L(t+1)e^{-\alpha t},\qquad \int_{|\xi|\ge\eta}\left|\frac{\partial}{\partial\xi}\hat{H}_5(\xi,t)\right|^2 d\xi\le L(t^2+1)e^{-\alpha t},$$

for all $t\ge 1$,

$$(1.32)\qquad \int_{|\xi|\le\rho}|\hat{H}_5(\xi,t)|\,d\xi\le Lt^{-3/2},\qquad \int_{|\xi|\le\rho}\left|\frac{\partial}{\partial\xi}\hat{H}_5(\xi,t)\right|^2 d\xi\le Lt^{-1/2}, \quad \text{for all } t\ge 1,$$

$$(1.33)\quad \int_{|\xi|\geq\eta}\left|\xi\hat{H}_5(\xi,t)\right|^2 d\xi\leq L(t^2+1+t^{-3/2})e^{-\alpha t},$$

$$\int_{|\xi|\geq\eta}\left|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_5(\xi,t)\right)\right|^2 d\xi\leq L(t^2+t^{-1/2})e^{-\alpha t}\quad \text{for all } t>0,$$

$$(1.34)\quad \int_{|\xi|\leq\rho}\left|\xi\hat{H}_5(\xi,t)\right|^2 d\xi\leq L(t+1)^{-7/2},\qquad \int_{|\xi|\leq\rho}\left|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_5(\xi,t)\right)\right|^2 d\xi\leq L(t+1)^{-3/2}$$

for all $t>0$.

From (1.27), (1.28), (1.29) and the identity

$$(1.35)\qquad \mathcal{F}^{-1}\left[e^{-\xi^2 t/2}\sin(\xi t)\right]=\frac{1}{2i}\frac{1}{\sqrt{2\pi t}}\left(e^{-(x+t)^2/2t}-e^{-(x-t)^2/2t}\right),$$

it follows that $G_{12}(x,t)\in C([0,\infty);L^1(R_x))$ and $\|G_{12}(x,t)\|\leq M_4$ for all $t\geq 0$. By (1.31) and (1.32), $\|\hat{H}_5(\xi,t)\|\leq Lt^{-3/2}$ and $\|(\partial/\partial\xi)\hat{H}_5(\xi,t)\|_{L^2}\leq Lt^{-1/4}$ for all $t\geq 1$. Thus, substituting H_5 for H_2 in (1.14) and taking $T=t^{5/6}$, it is shown that $H_5(x,t)\in C([1,\infty);L^1(R_x))$ and $\|H_5(x,t)\|\leq M_4 t^{-2/3}$ for all $t\geq 1$. Finally, with the aid of (1.33), (1.34) and (1.20) (substituting $(\partial/\partial x)H_5$ for H_2), we deduce that $(\partial/\partial x)H_5(x,t)\in C((0,\infty);L^1(R_x))$,

$$(1.36)\qquad \left\|\frac{\partial}{\partial x}H_5(x,t)\right\|\leq M_5 t^{-1/2}\quad \text{for small } t>0,$$

$$(1.37)\qquad \left\|\frac{\partial}{\partial x}H_5(x,t)\right\|\leq M_6 t^{-5/4}\quad \text{for large } t>0,$$

where M_5 and M_6 are positive constants. □

We now proceed to get estimates for $\|G_{22}(x,t)\|$, $\|(\partial/\partial x)G_{22}(x,t)\|$ and $\|(\partial^2/\partial x^2)G_{22}(x,t)\|$ by defining

$$\hat{H}_6(\xi,t)=\hat{G}_{22}-e^{-\xi^2 t},$$

$$\hat{H}_7(\xi,t)=\hat{G}_{22}-e^{-\xi^2 t/2}\cos(\xi t),$$

$$\hat{H}_8(\xi,t)=\xi^2\hat{G}_{22}+e^{-t}-\xi^2 e^{-\xi^2 t/2}\cos(\xi t)\quad \text{and}$$

$$H_i(x,t)=\mathcal{F}^{-1}\hat{H}_i(\xi,t),\qquad i=6,7,8.$$

LEMMA 1.3. *$H_6(x,t)\in C([0,2];L^1(R_x))$, $H_7(x,t)\in C([1,\infty);L^1(R_x))$, $(\partial/\partial x)H_6(x,t)\in C([0,2];L^1(R_x))$, $(\partial/\partial x)H_7(x,t)\in C([1,\infty);L^1(R_x))$ and $H_8(x,t)\in C([0,\infty);L^1(R_x))$. For some positive constant M_7,*

$$(1.38)\qquad \|H_7(x,t)\|\leq M_7 t^{-1/6}\quad \textit{for all } t\geq 1,$$

$$(1.39)\qquad \left\|\frac{\partial}{\partial x}H_7(x,t)\right\|\leq M_7 t^{-3/4}\quad \textit{for all } t\geq 1\quad \textit{and}$$

$$(1.40)\qquad \|H_8(x,t)\|\leq M_7(t+t^{7/6})^{-1}\quad \textit{for all } t>0.$$

Proof. The proof is quite similar to those of the preceding lemmas and follows from the following estimates:

$$(1.41)\quad \|\hat{H}_6(\xi,t)\|\leq L,\qquad \left\|\frac{\partial}{\partial\xi}\hat{H}_6(\xi,t)\right\|_{L^2}\leq L\quad \text{for all } t\in[0,2],$$

$$(1.42)\quad \|\xi\hat{H}_6(\xi,t)\|_{L^2}\leq L,\qquad \left\|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_6(\xi,t)\right)\right\|_{L^2}\leq L\quad \text{for all } t\in[0,2],$$

$$(1.43)\quad \|\hat{H}_7(\xi,t)\| \le L(t+1)e^{-\alpha t}+Lt^{-1},$$

$$\left\|\frac{\partial}{\partial\xi}\hat{H}_7(\xi,t)\right\|_{L^2} \le L(t+1)e^{-\alpha t}+Lt^{1/4} \quad \text{for all } t\in[1,\infty),$$

$$(1.44)\quad \|\xi\hat{H}_7(\xi,t)\|_{L^2} \le L(t+1)e^{-\alpha t}+Lt^{-5/4},$$

$$\left\|\frac{\partial}{\partial\xi}\left(\xi\hat{H}_7(\xi,t)\right)\right\|_{L^2} \le L(t+1)e^{-\alpha t}+Lt^{-1/4} \quad \text{for all } t\in[1,\infty),$$

$$(1.45)\quad \|\hat{H}_8(\xi,t)\| \le L(t+1+t^{-3/2})e^{-\alpha t}+L(t+1)^{-2},$$

$$\left\|\frac{\partial}{\partial\xi}\hat{H}_8(\xi,t)\right\|_{L^2} \le L(t+1+t^{-3/4})e^{-\alpha t}+L(t+1)^{-3/4} \quad \text{for all } t>0. \qquad \square$$

We conclude this section by recalling some elementary properties of functions of bounded variation which will be used in §2.

LEMMA 1.4. *Let f be a function such that $f\in C^1(R;R)$, $\dot{f}(\cdot)$ is locally Lipschitzian, $f(0)=0$ and $\dot{f}(0)=0$. Then for any $u,v\in L^1\cap \mathrm{BV}$, we have $f(u)$, $f(v)\in L^1\cap \mathrm{BV}$. In particular, for any given $M>0$, there exists a positive constant C_M such that $\|u_x\|$, $\|v_x\|\le M$ implies the following inequalities*:

$$(1.46)\quad \|f(u)\| \le C_M\|u_x\|\|u\|, \qquad \left\|\frac{d}{dx}f(u)\right\| \le C_M\|u_x\|^2,$$

$$(1.47)\quad \|f(u)-f(v)\| \le C_M\big(\|u_x\|+\|v_x\|\big)\|u-v\|,$$

$$\left\|\frac{d}{dx}f(u)-\frac{d}{dx}f(v)\right\| \le C_M\big(\|u_x\|+\|v_x\|\big)\|u_x-v_x\|.$$

Proof. Let δ_ε denote the Friedrichs mollifier. Since $u\in L^1\cap \mathrm{BV}$, $(u*\delta_\varepsilon)(x)\to u(x)$ as $\varepsilon\to 0$ for almost all x and $|u(x)|=|\int_{-\infty}^{x} d(du/dx)|\le \|u_x\|\le M<\infty$ for almost all x, where M is a positive constant. Therefore, $f(u*\delta_\varepsilon)\to f(u)$ in the sense of distributions. But $\|f(u*\delta_\varepsilon)\| \le C_M\|u*\delta_\varepsilon\|_{L^\infty}\|u*\delta_\varepsilon\| \le C_M\|u\|_{L^\infty}\|u\| \le C_M\|u_x\|\,\|u\|$ and $\|(d/dx)f(u*\delta_\varepsilon)\| = \|\dot{f}(u*\delta_\varepsilon)(u_x*\delta_\varepsilon)\| \le C_M\|u\|_{L^\infty}\|u_x\| \le C_M\|u_x\|^2$ hold for each $\varepsilon>0$, where C_M is the Lipschitz constant of $\dot{f}(\cdot)$ on the interval $[-M,M]$. These inequalities imply that $f(u)\in L^1\cap \mathrm{BV}$, $\|f(u)\|\le C_M\|u_x\|\,\|u\|$ and $\|(d/dx)f(u)\|\le C_M\|u_x\|^2$. The remaining assertions can be proved in a similar fashion. $\square$

LEMMA 1.5. *Under the same assumptions on f as in Lemma* 1.4, $f(u)\in C((0,\infty);L^1\cap \mathrm{BV})\cap C^1((0,\infty);\mathfrak{M})$ *for each* $u\in C((0,\infty);L^1\cap \mathrm{BV})\cap C^1((0,\infty);\mathfrak{M})$ *where $\mathfrak{M}$ is the Banach space of all finite measures.*

Proof. It is easy to see that $u*\delta_\varepsilon\in C((0,\infty);C^\infty\cap W^{1,1})\cap C^1((0,\infty);C^\infty\cap L^1)$, where the convolution is taken with respect to the x variable alone. Hence $(d/dt)f(u*\delta_\varepsilon)=\dot{f}(u*\delta_\varepsilon)(u_t*\delta_\varepsilon)$. Now fix any $0<\xi<\eta<\infty$. Since $\|u_x(t)\|$ is uniformly bounded on $[\xi,\eta]$, we can choose a suitable constant C independent of $\varepsilon>0$ such that

$$\left\|\frac{d}{dt}f(u*\delta_\varepsilon)(t)\right\| \le C\|u(t)\|_{L^\infty}\|u_t(t)\| \quad \text{for all } t\in[\xi,\eta],$$

$$\left\|\frac{d}{dt}f(u*\delta_\varepsilon)(t_1)-\frac{d}{dt}f(u*\delta_\varepsilon)(t_2)\right\|$$

$$\le C\|u(t_1)\|_{L^\infty}\|u_t(t_1)-u_t(t_2)\|+C\|u(t_1)-u(t_2)\|_{L^\infty}\|u_t(t_2)\|$$

$$\text{for all } t_1,t_2\in[\xi,\eta].$$

By the generalized Ascoli theorem (see [6]), it follows that the closure of $\{(d/dt)f(u*\delta_\varepsilon)\}_{\varepsilon>0}$ is compact in $C([\xi,\eta]; \tilde{\mathfrak{M}})$, where $\tilde{\mathfrak{M}}$ is the vector space $\mathfrak{M}$ equipped with the weak* topology. Since C_0, the space of all continuous functions vanishing at infinity, is separable, every weak* compact subset of $\tilde{\mathfrak{M}}$ is metrizable. Therefore, the closure of $\{(d/dt)f(u*\delta_\varepsilon)\}_{\varepsilon>0}$ is metrizable in the topology induced by $C([\xi,\eta];\tilde{\mathfrak{M}})$. Moreover, $f(u*\delta_\varepsilon)\to f(u)$ in the sense of distributions over $R\times(0,\infty)$, from which it follows that $(d/dt)f(u)=\lim_{\varepsilon\to 0}(d/dt)f(u*\delta_\varepsilon)\in C([\xi,\eta];\tilde{\mathfrak{M}})$. From the above inequality,

$$\|(d/dt)f(u)(t_1)-(d/dt)f(u)(t_2)\|\le C\|u(t_1)\|_{L^\infty}\|u_t(t_1)-u_t(t_2)\| + C\|u(t_1)-u(t_2)\|_{L^\infty}\|u_t(t_2)\|$$

holds for all $t_1, t_2\in[\xi,\eta]$. So we have shown $(d/dt)f(u)\in C([\xi,\eta];\mathfrak{M})$. Since ξ,η are arbitrary, $(d/dt)f(u)\in C((0,\infty);\mathfrak{M})$. The remainder of the proof is simpler and will thus be omitted. □

Remark 1.6. In order to study $\mathcal{F}^{-1}[e^{t\hat{A}(\xi)}]$, we have used here the explicit expression for $e^{t\hat{A}(\xi)}$. An alternative way would have been to analyze $\mathcal{F}^{-1}[e^{t\hat{A}(\xi)}]$ indirectly with the aid of asymptotic expansions and the Dunford integral, i.e., $e^{t\hat{A}(\xi)}=(1/2\pi i)\int_\Gamma e^{zt}[zI-\hat{A}(\xi)]^{-1}dz$. This approach may be useful in other similar but more complicated problems for which an explicit expression for $e^{t\hat{A}(\xi)}$ is not available.

2. Nonlinear problem. Equation (0.1) can be put in the form:

$$u_t=v_x,\qquad v_t=\dot\sigma(0)u_x+v_{xx}+[\sigma(u)-\dot\sigma(0)u]_x. \tag{2.1}$$

The basic properties of (0.4) established in §1 will now be used to solve (2.1) by means of the variation of constants formula. Without loss of generality, we normalize the function σ so that $\sigma(0)=0$ and $\dot\sigma(0)=1$. Now we state the main result:

THEOREM 2.1. *Suppose* $u_0(x), v_0(x)\in L^1\cap \mathrm{BV}$. *Then there is a small positive number* δ *such that* $\|u_0\|+\|u_{0x}\|+\|v_0\|+\|v_{0x}\|\le\delta$ *implies the existence of a unique global solution* $(u(t),v(t))$ *to* (2.1), (0.2) *in* $C([0,\infty);L^1\cap\mathrm{BV})\times[C([0,\infty);L^1)\cap C((0,\infty);W^{1,1})]$. *Furthermore,*

(i) $v_x(x,t)\to v_{0x}$, *as* $t\to 0$, *in the weak* topology of* $\mathfrak{M}$,

(ii) $v_t, v_{xx}\in C((0,\infty);\mathfrak{M})$,

(iii) $\|u(t)\|\le K$, $\|v(t)\|\le K$, $\|u_x(t)\|\le K(t+1)^{-1/2}$, $\|v_x(t)\|\le K(t+1)^{-1/2}$, $\|v_t(t)\|\le\tilde K t^{-1/2}$ *and* $\|v_{xx}(t)\|\le\tilde K t^{-1/2}$ *hold for all* $t>0$, *where* K *and* $\tilde K$ *are positive constants depending only on* δ.

Proof. Let us define the function space X as follows:

$$X=\left\{\begin{array}{l}(u(x,t),v(x,t)):\\ \quad u\in C([0,\infty);L^1\cap\mathrm{BV}),\quad v\in C([0,\infty);L^1)\cap C((0,\infty);W^{1,1}),\\ \quad u(x,0)=u_0,\quad v(x,0)=v_0,\quad \|u\|\le K,\quad \|v\|\le K,\\ \quad \|u_x\|\le K(t+1)^{-1/2},\quad \|v_x\|\le K(t+1)^{-1/2}\quad \text{for all } t\ge 0,\end{array}\right. \tag{2.2}$$

where $K>0$ is a constant which will be determined later and u_0, v_0 are given functions in $L^1\cap\mathrm{BV}$. X is a complete metric space equipped with the metric

$$d((u_1,v_1),(u_2,v_2))\overset{\text{def}}{=}\sup_{t\in[0,\infty)}\big(\|u_1-u_2\|+\|v_1-v_2\| +\sqrt{t+1}\,\|u_{1x}-u_{2x}\|+\sqrt{t+1}\,\|v_{1x}-v_{2x}\|\big).$$

We shall consider the mapping T on X, where, for each $(u,v)\in X$, $T(u,v)$ is the pair $(\tilde{u},\tilde{v})$ defined by

$$\tilde{u}=G_{11}*u_0+G_{12}*v_0+\int_0^t G_{12}(t-\tau)*[\sigma(u)-u]_x d\tau,$$

$$\tilde{v}=G_{21}*u_0+G_{22}*v_0+\int_0^t G_{22}(t-\tau)*[\sigma(u)-u]_x d\tau.$$

(The convolution is with respect to the x variable only.) We want to show that T is a contraction mapping on X into X.

By Lemmas 1.1 to 1.3, we deduce

$$G_{11}*u_0\in C([0,\infty);L^1\cap \mathrm{BV}),\qquad \|G_{11}*u_0\|\le\lambda\|u_0\| \tag{2.3}$$

$$\left\|\frac{\partial}{\partial x}G_{11}*u_0\right\|\le\lambda(t+1)^{-1/2}(\|u_0\|+\|u_{0x}\|)\quad\text{for all } t\ge 0,$$

where λ is a positive constant,

$$G_{12}*v_0\in C([0,\infty);W^{1,1}),\qquad \|G_{12}*v_0\|\le\lambda\|v_0\|, \tag{2.4}$$

$$\left\|\frac{\partial}{\partial x}G_{12}*v_0\right\|\le\lambda(t+1)^{-1/2}\|v_0\|\quad\text{for all } t\ge 0,$$

$$G_{21}*u_0\in C([0,\infty);W^{1,1}),\qquad \|G_{21}*u_0\|\le\lambda\|u_0\|, \tag{2.5}$$

$$\left\|\frac{\partial}{\partial x}G_{21}*u_0\right\|\le\lambda(t+1)^{-1/2}\|u_0\|\quad\text{for all } t\ge 0,$$

$$G_{22}*v_0\in C([0,\infty);L^1)\cap C((0,\infty);W^{1,1}),\qquad \|G_{22}*v_0\|\le\lambda\|v_0\|, \tag{2.6}$$

$$\left\|\frac{\partial}{\partial x}G_{22}*v_0\right\|\le\lambda(t+1)^{-1/2}(\|v_0\|+\|v_{0x}\|)\quad\text{for all } t\ge 0.$$

It follows from Lemma 1.4 that, assuming $K\le 1$,

$$\|\sigma(u_1)-\sigma(u_2)-(u_1-u_2)\|\le C(\|u_{1x}\|+\|u_{2x}\|)\|u_1-u_2\|, \tag{2.7}$$

$$\|\sigma(u_1)_x-\sigma(u_2)_x-(u_1-u_2)_x\|\le C(\|u_{1x}\|+\|u_{2x}\|)\|u_{1x}-u_{2x}\|, \tag{2.8}$$

where C is the upper bound of $|\ddot{\sigma}(\cdot)|$ on $[-1,1]$. Hence it is obvious that

$$p\overset{\text{def}}{=}\int_0^t G_{12}(t-\tau)*[\sigma(u)-u]_x d\tau\in C([0,\infty);L^1\cap \mathrm{BV}), \tag{2.9}$$

$$\|p\|\le\int_0^t\frac{M}{\sqrt{t-\tau+1}}\frac{CK^2}{\sqrt{\tau+1}}d\tau\le MK^2,$$

$$\left\|\frac{\partial p}{\partial x}\right\|\le\min\left\{\int_{t/2}^t\frac{M}{\sqrt{t-\tau+1}}\frac{CK^2}{\tau+1}d\tau+\int_0^{t/2}M\left(e^{-(t-\tau)}+\frac{1}{t-\tau}\right)\frac{CK^2}{\sqrt{\tau+1}}d\tau,\right.$$

$$\left.\int_0^t\frac{M}{\sqrt{t-\tau+1}}\frac{CK^2}{\tau+1}d\tau\right\}$$

$$\le M(t+1)^{-1/2}K^2\quad\text{for all } t\ge 0,\quad\text{and}$$

$$(2.10)\quad q \overset{\text{def}}{=} \int_0^t G_{22}(t-\tau) * [\sigma(u)-u]_x d\tau \in C([0,\infty); W^{1,1}),$$

$$\|q\| \le \int_0^t \frac{M}{\sqrt{t-\tau}} \frac{CK^2}{\sqrt{\tau+1}} d\tau \le MK^2,$$

$$\left\| \frac{\partial q}{\partial x} \right\| \le \min\left\{ \int_{t/2}^t \frac{M}{\sqrt{t-\tau}} \frac{CK^2}{\tau+1} d\tau + \int_0^{t/2} M\left(e^{-(t-\tau)} + \frac{1}{t-\tau} \right) \frac{CK^2}{\sqrt{\tau+1}} d\tau, \right.$$

$$\left. \int_0^t \frac{M}{\sqrt{t-\tau}} \frac{CK^2}{\tau+1} d\tau \right\}$$

$$\le M(t+1)^{-1/2} K^2 \quad \text{for all } t \ge 0,$$

where, from now on, M will denote positive generic constants. By (2.3) to (2.10), $\tilde{u} \in C([0,\infty); L^1 \cap \mathrm{BV})$ and $\tilde{v} \in C([0,\infty): L^1) \cap C((0,\infty); W^{1,1})$. Furthermore,

$$(2.11)\quad \|\tilde{u}\|, \|\tilde{v}\| \le K_1\left(\|u_0\| + \|v_0\| + K^2\right),$$

$$(2.12)\quad \|\tilde{u}_x\|, \|\tilde{v}_x\| \le K_1(t+1)^{-1/2}\left(\|u_0\| + \|v_0\| + \|u_{0x}\| + \|v_{0x}\| + K^2\right)$$

hold for all $t \ge 0$, where K_1 is a positive constant. So $(\tilde{u}, \tilde{v}) \in X$ if $K < \min(1/2K_1, 1)$, $0 < \delta < K/2K_1$ and $\|u_0\| + \|v_0\| + \|u_{0x}\| + \|v_{0x}\| \le \delta$. Next we shall show that T is a contraction.

Let $(\tilde{u}_1, \tilde{v}_1) = T(u_1, v_1)$ and $(\tilde{u}_2, \tilde{v}_2) = T(u_2, v_2)$. Then

$$(2.13)\quad \|\tilde{u}_1 - \tilde{u}_2\| \le \int_0^t \frac{M}{\sqrt{t-\tau+1}} \frac{2CK}{\sqrt{\tau+1}} d((u_1,v_1),(u_2,v_2))\, d\tau$$

$$\le K_2 K d((u_1,v_1),(u_2,v_2)),$$

$$(2.14)\quad \|\tilde{u}_{1x} - \tilde{u}_{2x}\| \le d((u_1,v_1),(u_2,v_2))$$

$$\cdot \min\left\{ \int_{t/2}^t \frac{M}{\sqrt{t-\tau+1}} \frac{2CK}{\tau+1} d\tau \right.$$

$$\left. + \int_0^{t/2} M\left(e^{-(t-\tau)} + \frac{1}{t-\tau} \right) \frac{2CK}{\sqrt{\tau+1}} d\tau, \int_0^t \frac{M}{\sqrt{t-\tau+1}} \frac{2CK}{\sqrt{\tau+1}} d\tau \right\}$$

$$\le K_2 K (t+1)^{-1/2} d((u_1,v_1),(u_2,v_2))$$

hold for all $t \ge 0$ by (2.7) and (2.8), where K_2 is a positive constant. In the same way,

$$(2.15)\quad \|\tilde{v}_1 - \tilde{v}_2\| \le K_2 K d((u_1,v_1),(u_2,v_2)),$$

$$(2.16)\quad \|\tilde{v}_{1x} - \tilde{v}_{2x}\| \le K_2 K (t+1)^{-1/2} d((u_1,v_1),(u_2,v_2))$$

hold for all $t \ge 0$. Hence $d((\tilde{u}_1,\tilde{v}_1),(\tilde{u}_2,\tilde{v}_2)) \le 4K_2 K d((u_1,v_1),(u_2,v_2))$, for which it follows that if $K < \min(1/2K_1, 1, 1/4K_2, U)$, $0 < \delta < K/2K_1$ and $\|u_0\| + \|v_0\| + \|u_{0x}\| + \|v_{0x}\| \le \delta$, then T is a contraction mapping of X into X. Here the condition $K < U$ guarantees that for each $t \ge 0$, $u(x,t) > -U$ holds for almost all x.

Thus T has a unique fixed point which is the solution to (2.1), (0.2). The local solutions are unique in a larger space. Suppose (u_1, v_1) and (u_2, v_2) are two solutions to (2.1), (0.2) such that $u_1,\ u_2 \in C([0,T]; L^1) \cap L^\infty([0,T]\times R)$ and $v_1,\ v_2 \in C([0,T]; L^1)$. Then

$$u_1 - u_2 = \int_0^t \frac{\partial}{\partial x} G_{12}(t-\tau) * [\sigma(u_1) - u_1 - \sigma(u_2) + u_2]\, d\tau. \tag{2.17}$$

Thus,

$$\|u_1(t) - u_2(t)\| \le M \int_0^t \|u_1(\tau) - u_2(\tau)\|\, d\tau \quad \text{and} \tag{2.18}$$

$$\|v_1(t) - v_2(t)\| \le M \int_0^t \frac{1}{\sqrt{t-\tau}} \|u_1(\tau) - u_2(\tau)\|\, d\tau \tag{2.19}$$

hold for all $t \in [0,T]$, from which it follows that $u_1 \equiv u_2$ and $v_1 \equiv v_2$. Finally, we shall estimate $\|v_t\|$ and $\|v_{xx}\|$. It is obvious that

$$\frac{\partial}{\partial t} G_{12} = \frac{\partial}{\partial x} G_{22} \quad \text{and} \quad \frac{\partial}{\partial t} G_{22} = \frac{\partial}{\partial x} G_{12} + \frac{\partial^2}{\partial x^2} G_{22}.$$

Therefore,

$$\begin{aligned} \frac{\partial v}{\partial t}(t) = {} & \frac{\partial}{\partial x} G_{22}(t) * u_0 + \left(\frac{\partial}{\partial x} G_{12}(t) + \frac{\partial^2}{\partial x^2} G_{22}(t) \right) * v_0 + G_{22}\left(\frac{t}{2}\right) * [\sigma(u) - u]_x \left(\frac{t}{2}\right) \\ & + \int_0^{t/2} \left[\frac{\partial}{\partial x} G_{12}(t-\tau) + \frac{\partial^2}{\partial x^2} G_{22}(t-\tau) \right] * [\sigma(u) - u]_x(\tau)\, d\tau \\ & + \int_{t/2}^t \frac{\partial}{\partial x} G_{22}(t-\tau) * [\sigma(u) - u]_\tau(\tau)\, d\tau \quad \text{for all } t > 0. \end{aligned} \tag{2.20}$$

By Lemmas 1.2 to 1.5, we deduce that

$$\begin{aligned} \left\| \frac{\partial v}{\partial t} \right\| \le {} & M t^{-1/2} \|u_0\| + M(t+1)^{-1/2} \|v_0\| + M t^{-1/2} \|v_{0x}\| + M \left\| u_x\left(\frac{t}{2}\right) \right\|^2 \\ & + \int_0^{t/2} \{ M(t-\tau)^{-1} K^2 (\tau+1)^{-1/2} \\ & \qquad + M e^{-(t-\tau)} K^2 (\tau+1)^{-1} + M(t-\tau)^{-1} K^2 (\tau+1)^{-1} \}\, d\tau \\ & + \int_{t/2}^t M(t-\tau)^{-1/2} K^2 (\tau+1)^{-1} d\tau \\ \le {} & M_8 t^{-1/2} \quad \text{for all } t > 0, \end{aligned} \tag{2.21}$$

where M_8 is a positive constant depending only on δ. From $v_{xx} = v_t - \sigma(u)_x$ and the above inequality, it follows that $\|v_{xx}\| \le M_9 t^{-1/2}$ for all $t > 0$ for some positive constant M_9 depending only on δ. □

Acknowledgment. The author is very grateful to Professor C. Dafermos for his invaluable advice.

REFERENCES

[1] 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.

[2] C. M. DAFERMOS, *The mixed initial-boundary value problem for the equations of nonlinear one-dimensional viscoelasticity*, J. Differential Equations, 6 (1969), pp. 71–86.

[3] _____, *Conservation laws with dissipation*, in Nonlinear Phenomena in Mathematical Sciences, V. Lakshmikantham, ed., Academic Press, New York, 1981.

[4] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[5] J. M. GREENBERG, R. C. MACCAMY, V. J. MIZEL, *On the existence, uniqueness and stability of solutions of the equation* $\dot{\sigma}(u_x)u_{xx}+\lambda u_{xtx}=\rho_0 u_{tt}$, J. Math. Mech., 17 (1968), pp. 707–728.

[6] J. L. KELLY AND I. NAMIOKA, *Linear Topological Spaces*, Springer-Verlag, New York, 1963.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0005 $01.25/0

HOMOGENIZATION AND LINEAR THERMOELASTICITY*

GILLES A. FRANCFORT†

Abstract. We study homogenization of linear dynamic thermoelasticity with rapidly varying coefficients, using a semigroup approach. The resulting homogenized problem exhibits an unusual change in initial temperature.

A formal asymptotic analysis predicts fast time oscillations in the temperature field. These oscillations explain the temperature shift, and show that, for our problem, weak convergence in time is the best convergence that one can obtain.

Introduction. We discuss the problem of "homogenizing" the equations of linear thermoelasticity when the mechanical and thermal properties are periodic and rapidly varying. Following Bensoussan, Lions and Papanicolaou [1] and Sanchez-Palencia [7] and using a semigroup approach, we show rigorously that, as the period of the coefficients goes to zero, the solution of these equations converges to the solution of a related constant coefficient problem, the *homogenized* problem. Then using a formal multiple-scales method, we give what we believe to be a satisfying interpretation of some surprising features of the results.

Thermoelastic behavior is characterized by the coupling of hyperbolic equations of motion and a parabolic heat equation. This leads to several interesting phenomena in the homogenization process.

Fast time oscillations in the temperature field are observed; their phase is completely determined. Thus the solutions can only converge in a weak sense in time to a slowly varying homogenized solution.

Furthermore, the initial data for the homogenized problem are related to the initial data of the inhomogeneous problem by *a linear transformation* which is not a projection. We know of no other examples of such a phenomenon.

In §1, we formulate and prove the existence of a homogenized thermoelastic medium. Section 2 contains the more formal arguments and the fast oscillations results, which are at the root of the observed change in initial data.

1. Homogenization of the thermoelastic problem. To reduce the overwhelmingly cumbersome notations that characterize thermoelasticity, we will place ourselves in a scalar setting, that is, one where the displacement field is taken to be scalar valued. Duvaut and Lions [2] show, using Korn's theorem, that this is no loss of generality.

We consider a domain Ω of $\mathbb{R}^n$. The degree of smoothness of $\delta\Omega$ will depend on the type of boundary conditions adopted. We will always assume that $\delta\Omega$ is smooth enough for one to be in position to apply Rellich's theorem on compact imbeddings of Sobolev spaces (Folland [3, Chap. 6]).

We will refer to $Y=\Pi_{i=1}^{n}]0,y_i^0[$ as the "reference cell"; $|Y|$ is its volume.

*Received by the editors August 20, 1981 and in revised form March 11, 1982. This work was funded by the Office of Naval Research under grant ONR N00014-76-C0054, through the Department of Mechanical Engineering at Stanford University.

†Division of Applied Mechanics, Department of Mechanical Engineering, Stanford University, Stanford, California 94305.

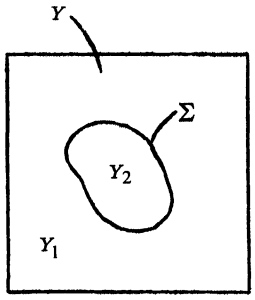


FIG 1.

If Σ is a smooth hypersurface dividing Y into Y_1 and Y_2 (see Fig. 1), we define $a_{ij}(y)$, $\lambda_{ij}(y)$, $\alpha_i(y)$, $\beta(y)$, $\rho(y)$ to be real Y-periodic functions, smooth and bounded on the closure of Y_1 and Y_2 but with Σ as potential surface of discontinuity.

Furthermore, $a_{ij}(y)$, $\lambda_{ij}(y)$ are assumed to be symmetric and strongly elliptic on Y, that is, there exists $\alpha>0$ such that for all ξ's in $\mathbb{R}^n$

$$a_{ij}(y)\ (\text{resp.}\lambda_{ij}(y))\ \xi_i\xi_j \geq \alpha\xi_i^2 \quad \text{on } Y, \tag{1.1}$$

$\beta(y)$ and $\rho(y)$ are positive and bounded away from zero. We finally choose α such that α^{-1} is a common upper bound of the L_∞-norms of the coefficients. We extend all coefficients to all of $\mathbb{R}^n$ by periodicity. Our equations are (Kupradze [5])

$$\begin{aligned}
&\rho\left(\frac{x}{\varepsilon}\right)\frac{\partial^2 u^\varepsilon}{\partial t^2}=\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\left(\frac{\partial u^\varepsilon}{\partial x_j}-\alpha_j\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\right)\right),\\
&\beta\left(\frac{x}{\varepsilon}\right)\frac{\partial \tau^\varepsilon}{\partial t}=\frac{\partial}{\partial x_i}\left(\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial \tau^\varepsilon}{\partial x_j}\right)-a_{ij}\left(\frac{x}{\varepsilon}\right)a_j\left(\frac{x}{\varepsilon}\right)\frac{\partial^2 u^\varepsilon}{\partial t\partial x_i}.
\end{aligned} \tag{1.2}$$

In (1.2), u^ε represents the displacement field and τ^ε the temperature increment field. The first equation is the scalar version of the equations of motion and the second is the heat conduction equation. The coupling between the equations results from consideration of the interaction between deformation and temperature: a temperature change induces strain and conversely. Finally, the rapid spatial oscillations in the coefficients translate the periodic structure of the body which comes from the assembling of ε-scaled versions of the reference cell Y. This body has to be thought of as made of a composite material where both constituents behave thermoelastically.

For the sake of simplicity we will only consider Dirichlet boundary conditions throughout:

$$u^\varepsilon=0, \qquad \tau^\varepsilon=0 \quad \text{on } \partial\Omega. \tag{1.3}$$

And for initial conditions, we will have:

$$u^\varepsilon(x,0)=f(x), \quad \frac{\partial u^\varepsilon}{\partial t}(x,0)=g(x), \quad \tau^\varepsilon(x,0)=k(x). \tag{1.4}$$

Our goal is to study the behavior of u^ε and τ^ε as ε, the period, goes to zero.

We define H to be:

$$H=H_0^1(\Omega)\times L_2(\Omega)\times L_2(\Omega). \tag{1.5}$$

On H, we define the operator A_ε:

(1.6)
$$A_\varepsilon = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{\rho\left(\frac{x}{\varepsilon}\right)} \frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial}{\partial x_j}\right) & 0 & -\frac{1}{\rho\left(\frac{x}{\varepsilon}\right)}\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\cdot\right) \\ 0 & -\frac{1}{\beta\left(\frac{x}{\varepsilon}\right)} a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial}{\partial x_i} & \frac{1}{\beta\left(\frac{x}{\varepsilon}\right)}\frac{\partial}{\partial x_i}\left(\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial}{\partial x_j}\right) \end{bmatrix}$$

with domain

(1.7) $$D(A_\varepsilon) = \{U=(u,u_t,\tau)\in H_0^1(\Omega)\times L_2(\Omega)\times H_0^1(\Omega) \text{ such that } A_\varepsilon U \text{ (taken in a distributional sense) belongs to } H\}.$$

Then the following proposition holds:

PROPOSITION 1.1. *A_ε generates in H a strongly continuous semigroup of operators $S_\varepsilon(t)$ such that*:

(1.8) $$\|S_\varepsilon(t)\| \le \alpha^{-1} \quad (\forall t>0).$$

Proof. We first consider for a fixed ε the norm

(1.9) $$|U|_\varepsilon^2 = \int_\Omega \left[a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial u}{\partial x_j}\frac{\partial \tilde{u}}{\partial x_i} + \rho\left(\frac{x}{\varepsilon}\right) u_t \tilde{u}_t + \beta\left(\frac{x}{\varepsilon}\right)\tau\tilde{\tau}\right] dx$$

where $\tilde{}$ denotes complex conjugation.

In view of the properties of the coefficients, $|\cdot|_\varepsilon$ is a norm on H, equivalent to the natural Sobolev norm on H, noted $\|\cdot\|$, that is, if U is in H,

(1.10) $$\alpha\|U\|^2 \le |U|_\varepsilon^2 \le \alpha^{-1}\|U\|^2.$$

In the norm $|\cdot|_\varepsilon$, A_ε generates a semigroup of contractions. Indeed, the domain $D(A_\varepsilon)$ is dense, since, though $\mathcal{C}_0^\infty(\Omega)$ functions do not belong to it, $\mathcal{C}_0^\infty(\Omega)$ functions whose conormal derivatives are 0 together with their third component on the only possible surfaces of discontinuity for the coefficients (i.e., the ε-scaled versions of Σ in each of the cells making up Ω) do belong to the domain $D(A_\varepsilon)$. Checking that A_ε is closed, that the range of $(1-A_\varepsilon)$ is H itself and that A_ε is dissipative offers no special difficulties (see Francfort [4] for full details). Note that the measure of the dissipation,

(1.11) $$\mathrm{Re}(A_\varepsilon U, U) = -\mathrm{Re}\left(\int_\Omega \lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial \tau}{\partial x_j}\frac{\partial \tilde{\tau}}{\partial x_i} dx\right) \le -\alpha|\nabla\tau|_{L_2(\Omega)}^2$$

(in view of the properties of the λ_{ij}'s), is precisely the physical dissipation due to the heat fluxes through the domain.

The result then follows from the application of Lumer–Phillips's theorem (Yosida [8, Chap. 9]). Therefore,

(1.12) $$|S_\varepsilon(t)U|_\varepsilon \le |U|_\varepsilon \quad \text{for any } U \text{ in } H,$$

and thus, using (1.10),

(1.13) $$\|S_\varepsilon(t)U\| \le \alpha^{-1}\|U\|,$$

which completes the proof. □

We now leave the time dependent formulation and examine the behavior of the resolvent of A_ε, $R_\lambda(A_\varepsilon)$ as ε goes to 0. At the end of this section we will reintroduce the time dependence by using some basic properties of semigroups.

It is a direct consequence of (1.8) (Yosida [8, Chap. 9]) that the right half complex plane belongs to the resolvent set of A_ε, for every ε. Let us consider $F=(f,g,k)$ to be an element of H. We take λ to be real strictly positive. The following string of equivalences holds:

(1.14)

$$
\begin{aligned}
&R_\lambda(A_\varepsilon)F=U_\varepsilon, \quad (U_\varepsilon=(u^\varepsilon,u_t^\varepsilon,\tau^\varepsilon))\\
&\quad\Leftrightarrow\quad \lambda u^\varepsilon-u_t^\varepsilon=f,\\
&\qquad \rho\left(\frac{x}{\varepsilon}\right)\lambda u_t^\varepsilon-\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\left(\frac{\partial u^\varepsilon}{\partial x_j}-\alpha_j\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\right)\right)=\rho\left(\frac{x}{\varepsilon}\right)g,\\
&\qquad \beta\left(\frac{x}{\varepsilon}\right)\lambda\tau^\varepsilon-\frac{\partial}{\partial x_i}\left(\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial\tau^\varepsilon}{\partial x_j}\right)+a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial u_t^\varepsilon}{\partial x_i}=\beta\left(\frac{x}{\varepsilon}\right)k,
\end{aligned}
$$

(1.15)

$$
\begin{aligned}
&\quad\Leftrightarrow\quad \lambda u^\varepsilon-u_t^\varepsilon=f,\\
&\qquad \lambda^2\rho\left(\frac{x}{\varepsilon}\right)u^\varepsilon-\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\left(\frac{\partial u^\varepsilon}{\partial x_j}-\alpha_j\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\right)\right)=\rho\left(\frac{x}{\varepsilon}\right)(\lambda f+g),\\
&\qquad \lambda\beta\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon-\frac{\partial}{\partial x_i}\left(\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial\tau^\varepsilon}{\partial x_j}\right)+\lambda a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial u^\varepsilon}{\partial x_i}\\
&\qquad\qquad =\beta\left(\frac{x}{\varepsilon}\right)k+a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial f}{\partial x_i}.
\end{aligned}
$$

The last two equations (1.15) have a unique solution $v^\varepsilon=\lambda u^\varepsilon$, τ^ε in $(H_0^1(\Omega))^2$, since the Dirichlet form d_ε defined as

$$
\begin{aligned}
d_\varepsilon((v^\varepsilon,\tau^\varepsilon),(\zeta,\eta))=&\frac{1}{\lambda}\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial v^\varepsilon}{\partial x_j}\frac{\partial\tilde\zeta}{\partial x_i}dx+\lambda\int_\Omega\rho\left(\frac{x}{\varepsilon}\right)v^\varepsilon\tilde\zeta\,dx\\
&-\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\frac{\partial\tilde\zeta}{\partial x_i}dx+\int_\Omega\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial\tau^\varepsilon}{\partial x_j}\frac{\partial\tilde\eta}{\partial x_i}dx\\
&+\lambda\int_\Omega\beta\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\tilde\eta\,dx+\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial v^\varepsilon}{\partial x_i}\tilde\eta\,dx
\end{aligned}
\tag{1.16}
$$

is strictly coercive on $(H_0^1(\Omega))^2$, in view of the properties of the coefficients.

If we manage to find a limit for u^ε, τ^ε as ε goes to zero, then going back up through the string (1.14) will enable us to obtain the limit of $R_\lambda(A_\varepsilon)F$.

Performing the limiting process in (1.15) is the task of the homogenization method. Rather than exposing all the details of the argument, we merely mention the different steps that were performed, underlining only the ones that are not standard. For further details the reader is to refer to Bensoussan, Lions and Papanicolaou [1, Chap. 1, esp. §§3, 9 and 13], or, for our problem, to Francfort [4].

Firstly, one shows that u_ε and τ_ε are bounded in $(H_0^1(\Omega))^2$, which immediately implies the existence of a weakly convergent subsequence in $(H_0^1(\Omega))^2$ converging to

(u,τ). Since we ultimately show that any convergent subsequence converges to the same limit, we do not distinguish between the sequence and subsequences of this sequence.

Then, defining

$$
\begin{aligned}
&\sigma_i^\varepsilon = a_{ij}\left(\frac{x}{\varepsilon}\right)\left(\frac{\partial u^\varepsilon}{\partial x_j} - \alpha_j\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon\right) && \text{the stress,}\\
&\kappa_i^\varepsilon = \lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial \tau^\varepsilon}{\partial x_j} && \text{the heat flux,}\\
&\nu^\varepsilon = a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial u^\varepsilon}{\partial x_i},
\end{aligned}
\tag{1.17}
$$

it is easy to conclude that these quantities converge weakly in $L_2(\Omega)$ to σ_i, κ_i, ν, which in turn satisfy:

$$
\bar{\rho}\lambda^2 u - \frac{\partial \sigma_i}{\partial x_i} = \bar{\rho}(\lambda f + g), \qquad \bar{\beta}\lambda\tau - \frac{\partial \kappa_i}{\partial x_i} + \lambda\nu = \bar{\beta}k + \overline{a_{ij}\alpha_j}\frac{\partial f}{\partial x_i},
\tag{1.18}
$$

where, from now on, $\bar{\ }$ will denote the Y-average $\frac{1}{|Y|}\int_Y dy$.

It remains to determine σ_i, κ_i, and ν. This is the core of homogenization. To this effect we define $\chi_k(y)$, $\Theta_k(y)$, $\Psi(y)$ to be the *unique periodic solutions*, up to a constant, in $H^1(Y)$ of:

$$
\begin{aligned}
&-\frac{\partial}{\partial y_i}\left(a_{ij}(y)\frac{\partial \chi_k}{\partial y_j}\right) = -\frac{\partial a_{ik}}{\partial y_i}(y),\\
&-\frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial \Theta_k}{\partial y_j}\right) = -\frac{\partial \lambda_{ik}}{\partial y_i}(y),\\
&-\frac{\partial}{\partial y_i}\left(a_{ij}(y)\frac{\partial \Psi}{\partial y_j}\right) = -\frac{\partial}{\partial y_i}\left(a_{ij}(y)\alpha_j(y)\right).
\end{aligned}
\tag{1.19}
$$

Ψ can be considered as nonstandard with respect to the "classical" case. The functions:

$$
w_k^\varepsilon = x_k - \varepsilon\chi_k\left(\frac{x}{\varepsilon}\right), \qquad z_k^\varepsilon = x_k - \varepsilon\Theta_k\left(\frac{x}{\varepsilon}\right)
\tag{1.20}
$$

satisfy:

$$
\begin{aligned}
&\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial w_k^\varepsilon}{\partial x_j}\frac{\partial \tilde{\omega}}{\partial x_i}dx = 0,\\
&\int_\Omega \lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial z_k^\varepsilon}{\partial x_j}\frac{\partial \tilde{\mu}}{\partial x_i}dx = 0 \qquad \text{for any } \omega,\mu \text{ in } H_0^1(\Omega).
\end{aligned}
\tag{1.21}
$$

Taking ω and μ to be $\mathcal{C}_0^\infty(\Omega)$ functions and making use of (1.16), (1.21), we have:

(1.22)

$$
\begin{aligned}
d_\varepsilon\left((\lambda u^\varepsilon,\tau^\varepsilon),(\omega w_k^\varepsilon,\mu z_k^\varepsilon)\right) &- \int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial w_k^\varepsilon}{\partial x_j}\frac{\partial(\tilde{\omega}u^\varepsilon)}{\partial x_i}dx - \int_\Omega \lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial z_k^\varepsilon}{\partial x_j}\frac{\partial(\tilde{\mu}\tau^\varepsilon)}{\partial x_i}dx\\
&= \int_\Omega \rho\left(\frac{x}{\varepsilon}\right)(\lambda f+g)\tilde{\omega}w_k^\varepsilon dx + \int_\Omega\left(\beta\left(\frac{x}{\varepsilon}\right)k + a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial f}{\partial x_i}\right)\tilde{\mu}z_k^\varepsilon dx.
\end{aligned}
$$

In (1.22), we have in essence subtracted from the variational formulation of (1.15) appropriate expressions equal to 0 in order to eliminate products of weak convergences.

It is then possible to go to the limit in (1.22) in a way identical to Bensoussan, Lions and Papanicolaou [1, Chap. 1, §3]. Upon our performing this limiting process, σ_i and κ_i come out to be:

$$
\begin{aligned}
\sigma_i &= \overline{\left(a_{ij}-a_{kj}\frac{\partial \chi_i}{\partial y_k}\right)}\frac{\partial u}{\partial x_j}-\overline{\left(a_{ij}\alpha_j-a_{kj}\alpha_j\frac{\partial \chi_i}{\partial y_k}\right)}\tau,\\
\kappa_i &= \overline{\left(\lambda_{ij}-\lambda_{kj}\frac{\partial \Theta_i}{\partial y_k}\right)}\frac{\partial \tau}{\partial x_j}.
\end{aligned}
\tag{1.23}
$$

Determining ν requires some additional effort and the use of Ψ. One defines g^ε to be:

$$
g^\varepsilon = 1+\varepsilon\Psi\left(\frac{x}{\varepsilon}\right);
\tag{1.24}
$$

then it satisfies, for any ω in $H_0^1(\Omega)$:

$$
\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial g^\varepsilon}{\partial x_j}\frac{\partial \tilde{\omega}}{\partial x_i}dx=\int_\Omega a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial \tilde{\omega}}{\partial x_i}dx.
\tag{1.25}
$$

Repeating the procedure of (1.22) but with μ equal to 0 and w_k^ε replaced by g^ε, we determine ν to be:

$$
\nu=\overline{\left(a_{ij}\alpha_j-a_{ij}\frac{\partial \Psi}{\partial y_j}\right)}\frac{\partial u}{\partial x_i}+\overline{\left(a_{kj}\alpha_j\frac{\partial \Psi}{\partial y_k}\right)}\tau.
\tag{1.26}
$$

Defining a_{ij}, A_i, B_i, λ_{ij}, γ_i, σ to be

$$
\begin{aligned}
a_{ij} &= \overline{a_{ij}-a_{kj}\frac{\partial \chi_i}{\partial y_k}}, & \lambda_{ij} &= \overline{\lambda_{ij}-\lambda_{kj}\frac{\partial \Theta_i}{\partial y_k}},\\
A_i &= \overline{a_{ij}\alpha_j-a_{kj}\alpha_j\frac{\partial \chi_i}{\partial y_k}}, & \gamma_i &= \overline{a_{ij}\alpha_j}-A_i,\\
B_i &= \overline{a_{ij}\alpha_j-a_{ij}\frac{\partial \Psi}{\partial y_j}}, & \sigma &= \overline{a_{kj}\alpha_j\frac{\partial \Psi}{\partial y_k}},
\end{aligned}
\tag{1.27}
$$

it can be shown, using (1.19), that a_{ij} and λ_{ij} are symmetric positive definite hence invertible, that A_i and B_i are equal and that σ is positive.

We set:

$$
\alpha_i=a_{ik}^{-1}A_k=a_{ik}^{-1}B_k.
\tag{1.28}
$$

Recalling (1.18), (1.23), (1.26)–(1.28) yields:

$$
\begin{aligned}
&\bar{\rho}\lambda^2 u-a_{ij}\left(\frac{\partial^2 u}{\partial x_i\partial x_j}-\alpha_j\frac{\partial \tau}{\partial x_i}\right)=\bar{\rho}(\lambda f+g),\\
&(\bar{\beta}+\sigma)\lambda\tau-\lambda_{ij}\frac{\partial^2 \tau}{\partial x_i\partial x_j}+\lambda a_{ij}\alpha_j\frac{\partial u}{\partial x_i}=\bar{\beta}k+\overline{a_{ij}(y)\alpha_j(y)}\frac{\partial f}{\partial x_i},
\end{aligned}
\tag{1.29}
$$

and, in view of the properties of the a_{ij}'s and λ_{ij}'s, the Dirichlet form associated with (1.29) is strictly coercive on $(H_0^1(\Omega))^2$, hence (1.29) admits a unique solution in $(H_0^1(\Omega))^2$.

Then, using (1.14), we obtain the following proposition:

PROPOSITION 1.2. *$R_\lambda(A_\varepsilon)F$ converges weakly in $(H_0^1(\Omega))^3$ to the unique solution in $(H_0^1(\Omega))^3$ of*:

$$
\begin{aligned}
&\lambda u - u_t = f,\\
&\lambda \bar{\rho} u_t - a_{ij}\left(\frac{\partial^2 u}{\partial x_i \partial x_j} - \alpha_j \frac{\partial \tau}{\partial x_i}\right) = \bar{\rho} g,\\
&\lambda(\bar{\beta}+\sigma)\tau - \lambda_{ij}\frac{\partial^2 \tau}{\partial x_i \partial x_j} + a_{ij}\alpha_j \frac{\partial u_t}{\partial x_i} = \bar{\beta} k + \gamma_i \frac{\partial f}{\partial x_i}.
\end{aligned}
\tag{1.30}
$$

We then define A to be:

$$
A = \begin{pmatrix}
0 & 1 & 0\\
\dfrac{1}{\bar{\rho}} a_{ij}\dfrac{\partial^2}{\partial x_i \partial x_j} & 0 & -\dfrac{1}{\bar{\rho}} a_{ij}\alpha_j \dfrac{\partial}{\partial x_i}\\
0 & -\dfrac{1}{\bar{\beta}+\sigma} a_{ij}\alpha_j \dfrac{\partial}{\partial x_i} & \dfrac{1}{\bar{\beta}+\sigma}\lambda_{ij}\dfrac{\partial^2}{\partial x_i \partial x_j}
\end{pmatrix}.
\tag{1.31}
$$

It is simply a matter of reproducing the proof of Proposition 1.1, but with constant coefficients this time, to show that *A generates a semigroup of operators $S(t)$* such that

$$
\|S(t)\| \le \alpha' \quad \text{for any } t \ge 0. \tag{1.32}
$$

Renaming α^{-1} the maximum of α' and α^{-1}, we deduce from Proposition 1.2 and (1.32) the following corollary:

COROLLARY 1.2. *$R_\lambda(A_\varepsilon)F$ converges weakly in $(H_0^1(\Omega))^3$ to $R_\lambda(A)\underset{\sim}{F}$ where*:

$$
\underset{\sim}{F} = \begin{pmatrix} f, g, \dfrac{\bar{\beta} k + \gamma_i \dfrac{\partial f}{\partial x_i}}{\bar{\beta}+\sigma} \end{pmatrix}. \tag{1.33}
$$

Now, (1.8) implies that, for any U, there is a bounded subsequence of $S_\varepsilon(t)U$ that converges weak* in $L_\infty(\mathbb{R}_+, H)$ to $\mathcal{G}(t)$ an element of $L_\infty(\mathbb{R}_+, H)$. This is a direct consequence of the separability of $L_1(\mathbb{R}_+, H)$ and of Banach–Alaoglu's theorem (Rudin [8, Chap. 3]). Still identifying a sequence with its subsequences, we get that, for any V in H,

$$
\int_0^\infty e^{-\lambda t}\left(S_\varepsilon(t)U, V\right)_H dt \overset{\varepsilon \to 0}{\to} \int_0^\infty e^{-\lambda t}\left(\mathcal{G}_{(t)}, V\right)_H dt \tag{1.34}
$$

where $(\cdot,\cdot)_H$ is the natural inner product on H. But the resolvent of the generator of a semigroup applied on a vector U is equal to the Laplace transform of the semigroup acting on U (Yosida [8, Chap. 9]), thus:

$$
\int_0^\infty e^{-\lambda t}\left(S_\varepsilon(t)U, V\right)_H dt = \left(R_\lambda(A_\varepsilon)U, V\right)_H, \tag{1.35}
$$

which itself converges to

$$(R_\lambda(A)\underset{\sim}{U},V)_H=\int_0^\infty e^{-\lambda t}(S(t)\underset{\sim}{U},V)_H dt. \tag{1.36}$$

Since V is arbitrary, we finally obtain, using the uniqueness of Laplace transforms of scalar functions:

$$\mathcal{G}_{(t)}=S(t)\underset{\sim}{U} \qquad (t\geq 0). \tag{1.37}$$

We have proved in this section the following theorem:

THEOREM. *The generalized solution of* (1.2) *with Dirichlet boundary conditions and initial conditions* (f,g,k) *in* H *converges weak* in* $L_\infty(\mathbb{R}_+,H)$ *to the generalized solution of*:

$$\begin{aligned}\bar{\rho}\frac{\partial^2 u}{\partial t^2}&=a_{ij}\left(\frac{\partial^2 u}{\partial x_i\partial x_j}-\alpha_j\frac{\partial\tau}{\partial x_i}\right),\\ (\bar{\beta}+\sigma)\frac{\partial\tau}{\partial t}&=\lambda_{ij}\frac{\partial^2\tau}{\partial x_i\partial x_j}-a_{ij}\alpha_j\frac{\partial^2 u}{\partial t\partial x_i}\end{aligned} \tag{1.38}$$

with Dirichlet boundary conditions and initial conditions

$$\left(f,g,\frac{\bar{\beta}k+\gamma_i\dfrac{\partial f}{\partial x_i}}{\bar{\beta}+\sigma}\right). \tag{1.39}$$

Before concluding this section, let us emphasize once more the rather unusual change in initial temperature in (1.39).

2. Fast oscillations of the temperature field. Since, through a L_∞ weak* type of convergence, a rapidly oscillating function (like $e^{it/\varepsilon}$) goes to 0, it is fairly natural to expect a t/ε dependence of u^ε and τ^ε. This kind of problem is most easily addressed using asymptotic expansion techniques. We have already mentioned the semiheuristic character of this section, so that we will not dwell on the restrictions to the problem that would make the argument totally rigorous.

Recalling (1.2) we now suppose that u^ε and τ^ε are functions of *both* t and $\delta=t/\varepsilon$; ∂_t becomes $\partial_t+\frac{1}{\varepsilon}\partial_\delta$. We then Laplace transform (1.2) with respect to *both* t and δ, the dual variables being respectively ζ and μ. From now on:

$\hat{}$ will denote the t-Laplace transform,

$\tilde{}$ will denote the δ-Laplace transform,

$\check{}$ will denote $\hat{\tilde{}}$ or $\tilde{\hat{}}$.

In order to be able to perform these transformations, we need to impose initial conditions on both t and δ. We will set:

$$\begin{aligned}&u^\varepsilon(x;0,\delta)=f(x), && u^\varepsilon(x;t,0)=p(x,t),\\ &\frac{\partial u^\varepsilon}{\partial t}(x;0,\delta)=g(x), && \frac{\partial u^\varepsilon}{\partial\delta}(x;t,0)=q(x,t),\\ &\tau^\varepsilon(x;0,\delta)=k(x), && \tau^\varepsilon(x;t,0)=\Theta(x,t),\end{aligned} \tag{2.1}$$

where f, g, k are as before and p, q, Θ are unknown. We obtain:

(2.2)

$$\rho\left(\frac{x}{\varepsilon}\right)\left\{\left(\zeta^2\check{u}^\varepsilon-\frac{\zeta f}{\mu}-\frac{g}{\mu}\right)+\frac{2}{\varepsilon}(\zeta\mu\check{u}^\varepsilon-\zeta\hat{p})+\frac{1}{\varepsilon^2}(\mu^2\check{u}^\varepsilon-\mu\hat{p}-\hat{q})\right\}$$
$$=\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\left(\frac{\partial\check{u}^\varepsilon}{\partial x_j}-\alpha_j\left(\frac{x}{\varepsilon}\right)\check{\tau}^\varepsilon\right)\right),$$
$$\beta\left(\frac{x}{\varepsilon}\right)\left\{\left(\zeta\check{\tau}^\varepsilon-\frac{k}{\mu}\right)+\frac{1}{\varepsilon}(\mu\check{\tau}^\varepsilon-\hat{\Theta})\right\}$$
$$=\frac{\partial}{\partial x_i}\left(\lambda_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial\check{\tau}^\varepsilon}{\partial x_j}\right)-a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\left\{\frac{\partial}{\partial x_i}\left(\zeta\check{u}^\varepsilon-\frac{f}{\mu}\right)+\frac{1}{\varepsilon}\frac{\partial}{\partial x_i}(\mu\check{u}^\varepsilon-\hat{p})\right\}.$$

We seek an expansion of u^ε and τ^ε in the form

$$u^\varepsilon=\sum\varepsilon^i u_i(x,y,t,\delta),\qquad \tau^\varepsilon=\sum\varepsilon^i\tau_i(x,y,t,\delta)\quad\text{where } y=\frac{x}{\varepsilon}. \tag{2.3}$$

The dependence of the u_i's and τ_i's on y is taken to be *Y-periodic*. This is always what is assumed when performing double scaling in space in problems related to homogenization.

We also need to control the fast time behavior of u_i and τ_i. Since we would like them to be oscillating in δ, or, at least, to be such that

$$\lim_{T\to+\infty}\frac{1}{T}\int_0^T u_i(x,t,y,\delta)\,d\delta \text{ (respectively } \tau_i) \tag{2.4}$$

exist and be finite, we are led through Wiener's Tauberian theorem (Rudin [6, Chap. 9]) to suppose that

$$\lim_{\mu\to0}\mu\check{u}_i \text{ (respectively } \mu\check{\tau}_i) \text{ exists and is finite,} \tag{2.5}$$

and we will furthermore assume that this limit is to be taken pointwise in x and weakly in $H^1(Y)$ with regard to the y dependence.

With these considerations in mind we can proceed to replace $\partial/\partial x_i$ by $\partial/\partial x_i+\frac{1}{\varepsilon}\partial/\partial y_i$ and u^ε and τ^ε by their expansions in (2.2).

We obtain two "series" in ascending powers of ε starting at ε^{-2}; we successively identify the factors of each of these powers to 0. As factor of ε^{-2} we get:

$$\rho(y)(\mu^2\check{u}_0-\mu\hat{p}-\hat{q})=\frac{\partial}{\partial y_i}\left(a_{ij}(y)\frac{\partial\check{u}_0}{\partial y_j}\right),$$
$$\frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial\check{\tau}_0}{\partial y_j}\right)-a_{ij}(y)\alpha_j(y)\frac{\partial}{\partial y_i}(\mu\check{u}_0-\hat{p})=0. \tag{2.6}$$

Since the Dirichlet form associated to the operator

$$D=\rho(y)\mu^2\cdot-\frac{\partial}{\partial y_i}\left(a_{ij}(y)\frac{\partial\cdot}{\partial y_j}\right) \tag{2.7}$$

is strictly coercive on the subspace of $H^1(Y)$ consisting of Y-periodic functions, the first equation of (2.6) has a unique solution there; thus $\hat{p}/\mu+\hat{q}/\mu^2$ is the solution. Hence

$$\check{u}_0=\frac{\hat{p}}{\mu}+\frac{\hat{q}}{\mu^2}. \tag{2.8}$$

But, in view of (2.5), (2.8) implies that $\hat{q}=0$, thus $\check{u}_0$ is equal to $\hat{p}/\mu$ and *does not depend on* y. Inverting (2.8) we obtain that u_0 *does not depend on* δ *either*;

$$u_0(x,t)=p(x,t). \tag{2.9}$$

Then, from the second equation of (2.6),

$$\check{\tau}_0=\check{\tau}_0(x,\mu), \tag{2.10}$$

since the only periodic solution of that equation is a constant with respect to y.

As factor of ε^{-1} we get, using (2.8), (2.10):

$$\begin{aligned} &D\check{u}_1=\frac{\partial a_{ij}}{\partial y_i}(y)\frac{\partial \check{u}_0}{\partial x_j}-\frac{\partial}{\partial y_i}\big(a_{ij}(y)\alpha_j(y)\big)\check{\tau}_0,\\ &\beta(y)\big(\mu\check{\tau}_0-\hat{\Theta}\big)=\frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial \check{\tau}_1}{\partial y_j}\right)+\frac{\partial \lambda_{ij}}{\partial y_i}(y)\frac{\partial \check{\tau}_0}{\partial x_j}-\mu a_{ij}(y)\alpha_j(y)\frac{\partial \check{u}_1}{\partial y_i}. \end{aligned} \tag{2.11}$$

Defining χ_i^μ and Ψ^μ to be the unique periodic solutions in $H^1(Y)$ of

$$\begin{aligned} &\mu^2\chi_i^\mu-\frac{\partial}{\partial y_k}\left(a_{kj}(y)\frac{\partial \chi_i^\mu}{\partial y_j}\right)=-\frac{\partial a_{ki}}{\partial y_k}(y),\\ &\mu^2\Psi^\mu-\frac{\partial}{\partial y_k}\left(a_{kj}(y)\frac{\partial \Psi^\mu}{\partial y_j}\right)=-\frac{\partial}{\partial y_k}\big(a_{kj}(y)\alpha_j(y)\big), \end{aligned} \tag{2.12}$$

we obtain from the first equation of (2.11):

$$\check{u}_1=-\chi_j^\mu\frac{\partial \check{u}_0}{\partial x_j}+\Psi^\mu\check{\tau}_0. \tag{2.13}$$

Then, integrating the second equation of (2.11) with respect to y and defining γ_i^μ and σ^μ to be the analogues of γ_i and σ for χ_i^μ and Ψ^μ as in (1.27),

$$\check{\tau}_0=\frac{1}{\mu}\frac{\overline{\beta}\hat{\Theta}+\gamma_i^\mu\dfrac{\partial \hat{p}}{\partial x_i}}{\overline{\beta}+\sigma^\mu}\overset{\text{(def.)}}{=}\frac{\hat{\eta}^\mu}{\mu} \tag{2.14}$$

where $\overline{\ }$ denotes the Y-average $\int_Y dy$. We introduce Λ_k^μ and H^μ to be the unique periodic solutions in $H^1(Y)$, up to a constant, of:

$$\begin{aligned} &-\frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial \Lambda_k^\mu}{\partial y_j}\right)=-\left(a_{ij}(y)\alpha_j(y)\frac{\partial \chi_k^\mu}{\partial y_i}-\frac{\beta(y)}{\overline{\beta}}\gamma_k^\mu\right),\\ &-\frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial H^\mu}{\partial y_j}\right)=-\left(a_{ij}(y)\alpha_j(y)\frac{\partial \Psi^\mu}{\partial y_i}-\frac{\beta(y)}{\overline{\beta}}\sigma^\mu\right). \end{aligned} \tag{2.15}$$

The equations (2.15) are well posed since the Y-averages of the right-hand sides do vanish by definition of γ_i^μ and σ^μ.

Recalling $\Theta_j(y)$, we obtain for $\check{\tau}_1$:

$$\check{\tau}_1 = -\frac{1}{\mu}\Theta_j(y)\frac{\partial \hat{\eta}^\mu}{\partial x_j} - \Lambda_j^\mu \frac{\partial \hat{p}}{\partial x_j} + H^\mu \hat{\eta}_\mu + \text{an arbitrary function of } x \text{ only.} \tag{2.16}$$

Finally, as factor of ε^0, we get:

(2.17)

$$\begin{aligned}
&\rho(y)\left(\zeta^2 \check{u}_0 - \zeta\frac{f}{\mu} - \frac{g}{\mu} + 2\zeta\mu\check{u}_1\right) + D\check{u}_2 \\
&\quad = \frac{\partial}{\partial y_i}\left(a_{ij}(y)\frac{\partial \check{u}_1}{\partial x_j}\right) + a_{ij}(y)\frac{\partial^2 \check{u}_0}{\partial x_i \partial x_j} + a_{ij}(y)\frac{\partial^2 \check{u}_1}{\partial x_i \partial y_j} \\
&\qquad - a_{ij}(y)\alpha_j(y)\frac{\partial \check{\tau}_0}{\partial x_i} - \frac{\partial}{\partial y_i}\left(a_{ij}(y)\alpha_j(y)\check{\tau}_1\right), \\
&\beta(y)\left(\zeta\check{\tau}_0 - \frac{k}{\mu} + \mu\check{\tau}_1\right) = \frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial \check{\tau}_2}{\partial y_j}\right) + \frac{\partial}{\partial y_i}\left(\lambda_{ij}(y)\frac{\partial \check{\tau}_1}{\partial x_j}\right) \\
&\qquad + \lambda_{ij}(y)\frac{\partial^2 \check{\tau}_1}{\partial x_i \partial y_j} + \lambda_{ij}(y)\frac{\partial^2 \check{\tau}_0}{\partial x_i \partial x_j} \\
&\qquad - a_{ij}(y)\alpha_j(y)\frac{\partial}{\partial x_i}\left(\zeta\check{u}_0 - \frac{f}{\mu}\right) - a_{ij}(y)\alpha_j(y)\frac{\partial}{\partial y_i}(\zeta\check{u}_1) \\
&\qquad - a_{ij}(y)\alpha_j(y)\frac{\partial}{\partial x_i}(\mu\check{u}_1) - a_{ij}(y)\alpha_j(y)\frac{\partial}{\partial y_i}(\mu\check{u}_2).
\end{aligned}$$

We integrate both equations of (2.17) with respect to y; making use of all the previous results of this section, we obtain:

$$\begin{aligned}
&\bar{\rho}(\zeta^2\hat{p} - \zeta f - g) + \mu^3\overline{\rho\check{u}_2} = a_{ij}^\mu \frac{\partial^2 \hat{p}}{\partial x_i \partial x_j} - a_{ij}^\mu \alpha_j^\mu \frac{\partial \hat{\eta}^\mu}{\partial x_i}, \\
&\bar{\beta}\,(\zeta\hat{\eta}^\mu - k) + \mu^2\overline{\beta\check{\tau}_1} \\
&\qquad = \lambda_{ij}\frac{\partial \hat{\eta}^\mu}{\partial x_i \partial x_j} - \mu\overline{\lambda_{ij}(y)\frac{\partial \Lambda_k^\mu}{\partial y_j}}\frac{\partial^2 \hat{p}}{\partial x_i \partial x_k} + \mu\overline{\lambda_{ij}(y)\frac{\partial H^\mu}{\partial y_j}}\frac{\partial \hat{\eta}_\mu}{\partial x_i} \\
&\qquad\quad - \zeta a_{ij}^\mu \alpha_j^\mu \frac{\partial \hat{p}}{\partial x_i} + \overline{a_{ij_{(y)}}\alpha_{j_{(y)}}}\frac{\partial f}{\partial x_i} - \zeta\sigma^\mu\hat{\eta}^\mu \\
&\qquad\quad - \mu^2\overline{a_{ij}(y)\alpha_j(y)\frac{\partial \check{u}_1}{\partial x_i}} - \mu^2\overline{a_{ij}(y)\alpha_j(y)\frac{\partial \check{u}_2}{\partial y_i}},
\end{aligned} \tag{2.18}$$

where a_{ij}^μ, α_j^μ are to χ_j^μ and Ψ^μ what a_{ij} and α_j are to χ_j and Ψ in (1.27).

We now consider the limit of (2.18) as μ goes to 0. The following result holds:

PROPOSITION 2.1. *χ_k^μ, Ψ^μ, $\mu\Lambda_k^\mu$, μH^μ go respectively to χ_i, Ψ, 0 and 0 strongly in $H^1(Y)/\mathbb{R}$ as μ goes to 0. Hence a_{ij}^μ, α_j^μ, γ_i^μ, σ^μ go to a_{ij}, α_j, γ_i, σ.*

The proof of this proposition, which involves some basic estimates in $H^1(Y)/\mathbb{R}$, will not be given here; refer to Francfort [4] for the details.

Proposition 2.1 together with (2.5) enables us to perform the limiting process. Upon doing so, we come up with a set of two equations for $\hat{p}$ and $\hat{\Theta}$ which, together with the limit of (2.14), can be interpreted in the time dependent domain. p and Θ

satisfy:

$$
\begin{aligned}
&\Theta(x,t)=\frac{(\bar{\beta}+\sigma)\eta(x,t)-\gamma_i\dfrac{\partial p}{\partial x_i}(x,t)}{\bar{\beta}},\\
&\bar{\rho}\,\frac{\partial^2 p}{\partial t^2}=a_{ij}\frac{\partial^2 p}{\partial x_i\partial x_j}-a_{ij}\alpha_j\frac{\partial \eta}{\partial x_i},\\
&(\bar{\beta}+\sigma)\frac{\partial \eta}{\partial t}=\lambda_{ij}\frac{\partial^2 \eta}{\partial x_i\partial x_j}-a_{ij}\alpha_j\frac{\partial^2 p}{\partial t\partial x_i},\\
&p(x,0)=f,\quad \frac{\partial p}{\partial t}(x,0)=g,\quad \eta(x,0)=\frac{\bar{\beta}k+\gamma_i\dfrac{\partial f}{\partial x_i}}{\bar{\beta}+\sigma},
\end{aligned}
\tag{2.19}
$$

where $\hat{\eta}$ is the limit of $\hat{\eta}^\mu$ as μ goes to 0.

It is clear that $\eta(x,t)$ can be identified with $\tau(x,t)$, the homogenized temperature field, and $p(x,t)$ with $u(x,t)$, the homogenized displacement field. Replacing Θ by its value in (2.14) we also obtain an expression for the leading term of the asymptotic expansion of τ^ε, that is τ_0; its δ-Laplace transform satisfies:

$$
\tilde{\tau}_0=\frac{1}{\mu}\left(\frac{\bar{\beta}+\sigma}{\bar{\beta}+\sigma^\mu}\eta+\frac{(\gamma_i^\mu-\gamma_i)}{\bar{\beta}+\sigma^\mu}\frac{\partial p}{\partial x_i}\right).
\tag{2.20}
$$

This expression is not explicitly invertible in general, in view of the complicated dependence of γ_i^μ and σ^μ on μ. It is, however, possible to show from (2.20) that τ_0 is the solution of a Volterra integral equation of the second kind (see Francfort [4]). Such an equation does not provide more information about τ_0 than (2.20) itself and it is therefore of little value for our purpose. The following proposition holds:

PROPOSITION 2.2. *σ^μ and γ_i^μ go to zero as μ goes to $+\infty$.*

The proof of this last proposition uses the same estimates as the ones that establish Proposition 2.1.

Propositions 2.1 and 2.2 enable us to conclude that, as μ goes to 0, $\mu\tilde{\tau}_0$ goes to η, whereas as μ goes to $+\infty$, $\mu\tilde{\tau}_0$ goes to Θ. In a time dependent context, these facts translate into statements on the behavior of τ_0 near infinity and near the origin,

$$
\begin{aligned}
&\lim_{\delta\to+\infty}\frac{1}{\delta}\int_0^\delta \tau_0(x,t,\delta')\,d\delta'=\tau(x,t),\\
&\lim_{\delta\to 0+}\frac{1}{\delta}\int_0^\delta \tau_0(x,t,\delta')\,d\delta'=\Theta(x,t),
\end{aligned}
\tag{2.21}
$$

provided these limits exist. The second equation of (2.21) is consistent with our self-imposed δ-initial conditions. The first equation shows that the fast oscillations of the leading term τ_0 of the asymptotic expansion of τ^ε are centered about $\tau(x,t)$, the solution of the homogenized problem. The initial condition $\tau(x,0)$ is the initial average of the oscillating function τ_0. This average is generally different from the initial value of τ_0 (or τ^ε). In other words, the shift in initial conditions is necessary if the initial "phase" is not zero. This contrasts with the method of geometrical optics in which the initial phase is arbitrary. It appears anyway that a geometrical optics type ansatz in place of (2.1) will fail since, if the solutions of (2.12) are sums of terms of more than one frequency in δ, the fast oscillations need not be periodic in δ.

Note that the first equation of (2.21) is the only specific information that we managed to obtain about the large time behavior of τ_0. We do not know, in a general case, for how long a time our expression of the fast oscillations remains valid.

Note also that fast oscillations do not appear in the leading term of the asymptotic expansion of u^ε, which could have been foreseen in view of the convergence obtained for u^ε and u_t^ε in §1.

Conclusion. Numerical computations corroborate the results of §§1 and 2 and confirm that fast oscillations are indeed the phenomenon leading to this unusual change in initial data [4]. It is the first time, to the author's knowledge, that fast oscillations in time are evidenced in a homogenization problem with time independent coefficients.

If seeking a more physical explanation, one could examine the entropy associated with the problem:

$$s=\beta\left(\frac{x}{\varepsilon}\right)\tau^\varepsilon+a_{ij}\left(\frac{x}{\varepsilon}\right)\alpha_j\left(\frac{x}{\varepsilon}\right)\frac{\partial u^\varepsilon}{\partial x_i}.$$

It is fairly straightforward, using the results of §2 and some of the steps performed there, to show that there is no fast dependence in time of the space average of the leading term in the expansion of s^ε. That the macroscopic entropy of this body is a *slowly varying quantity* appears to be a sound idea and does fit our physical intuition. A fast oscillation in the temperature field is the effect that balances the space oscillations of the strains due to the inhomogeneities of the coefficients and allows the entropy to evolve slowly at its own pace. In this respect the unusual initial change in temperature is needed to insure that no fast change in entropy is taking place at time zero.

To conclude this study, let us point out that choosing the entropy as the natural variable in place of the temperature introduces space derivatives of the third order and thereby prohibits a rigorous analysis of the type performed in §1. A perturbation analysis using double scaling is feasible but eventually leads to reintroducing the temperature field as the proper variable.

Acknowledgments. The author is extremely grateful to Professor G. Herrmann, his Ph.D. advisor.

He also wishes to express his profound gratitude to Professor R. Caflisch for his constant assistance throughout this study, and to Professor R. S. Phillips and Dr. J. Goodman for their invaluable suggestions.

REFERENCES

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

[2] G. DUVAUT AND J. L. LIONS, *Variational Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[3] G. FOLLAND, *Introduction to Partial Differential Equations*, Princeton Univ. Press, Princeton, NJ, 1976.

[4] G. FRANCFORT, Ph. D. Dissertation, Stanford University, April, 1982.

[5] V. D. KUPRADZE, *Three-Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity*, North-Holland, Amsterdam, 1979.

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

[7] E. SANCHEZ-PALENCIA, *Non Homogeneous Media and Vibration Theory*, Monographs in Physics, 127, Springer-Verlag, New York, 1980.

[8] K. YOSIDA, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0006 $01.25/0

OSCILLATION PROPERTIES OF SOLUTIONS OF SECOND ORDER ELLIPTIC EQUATIONS*

NORIO YOSHIDA†

Abstract. Elliptic differential equations with variable coefficients are studied and sufficient conditions are derived for all solutions to be oscillatory in exterior domains in Euclidean n-space. Both superlinear and sublinear equations are considered.

1. Introduction. Oscillation theory for elliptic differential operators with variable coefficients has been developed by several authors; see e.g. Allegretto [1], [2], Noussair and Swanson [13], Swanson [18] and the references contained therein. We are concerned with the oscillatory behavior of solutions of semilinear uniformly elliptic equations of the form

$$(*) \qquad L[u] \equiv \sum_{i,j=1}^{n} D_i\big(a_{ij}(x)D_j u\big) + c(x,u) = f(x), \qquad x \in \Omega,$$

where Ω is an exterior domain in $\mathbf{R}^n$, i.e., Ω contains the complement of some n-ball in $\mathbf{R}^n$. In [12, p. 79] Noussair and Swanson posed the problem of establishing superlinear oscillation criteria for L in the case of variable coefficients $a_{ij}(x)$. An analogous open question in the sublinear case is mentioned in the paper of Kitamura and Kusano [7, p. 174]. The purpose of this paper is to obtain sufficient conditions for every solution of $(*)$ to be oscillatory in Ω. Our method is an adaptation of that used in [9], [12] and the key tool is similar to that used by Levine and Payne [10] and Suleĭmanov [17] in studying the nonexistence of entire solutions of nonlinear elliptic equations.

The superlinear results in §3 are obtained by using a fundamental lemma given in §2. In §4 we consider sublinear equations in the case of variable coefficients $a_{ij}(x)$.

Points in $\mathbf{R}^n$ will be denoted by $x=(x_1,\cdots,x_n)$, and differentiation with respect to x_i by D_i, $i=1,\cdots,n$. The notation $|x|$ will be used for the Euclidean length of $x \in \mathbf{R}^n$. We assume that there exists a positive number a such that $\mathbf{R}^n(a) \subset \Omega$, where $\mathbf{R}^n(a) \equiv \{x \in \mathbf{R}^n : |x| > a\}$ $(a>0)$. The domain $\mathfrak{D}_L(\Omega)$ of L is defined to be the set of all real-valued functions of class $C^2(\Omega)$. We assume that the following conditions hold throughout this paper:

(A-I) $c(x,\xi)$ is a real-valued continuous function in $\Omega \times \mathbf{R}^1$.

(A-II) $f(x)$ is a real-valued continuous function in Ω.

(A-III) $D_k D_l a_{ij}(x)$ are locally uniformly Hölder-continuous in Ω $(i,j,k,l=1,\cdots,n)$ and the matrix $A(x) \equiv (a_{ij}(x))_{i,j=1}^n$ is symmetric in Ω.

(A-IV) L is uniformly elliptic, i.e., for some positive constant μ (<1), the inequality

$$\mu|\xi|^2 \le \xi^T A(x)\xi \le \mu^{-1}|\xi|^2$$

holds for all $\xi \in \mathbf{R}^n$ and every point $x \in \Omega$, where the superscript T denotes the transpose.

DEFINITION. A function $u \in \mathfrak{D}_L(\Omega)$ is said to be *oscillatory* in Ω if it has arbitrarily large zeros, i.e., the set $\{x \in \Omega : u(x)=0\}$ is unbounded.

* Received by the editors February 23, 1982, and in revised form August 2, 1982.

† Department of Mathematics, Faculty of Engineering, Iwate University, Morioka, Japan.

2. Two lemmas. Let x_0 be a fixed point in $\mathbf{R}^n(a)$. From (A-III) and (A-IV) we conclude that there exists a fundamental solution $E(x)\in C^2(\mathbf{R}^n(a)\setminus\{x_0\})$ of the operator $P\equiv\sum_{i,j=1}^n D_i a_{ij}(x)D_j$ with a singularity at the point x_0, i.e.,

$$P[E(x)]=-\delta(x-x_0) \quad (\delta\text{: Dirac function})$$

(see [4, p. 84]). In agreement with what obtains for the Laplace operator Δ, we assume that

$$\sum_{i,j=1}^n a_{ij}(x)\big(D_iE(x)\big)\big(D_jE(x)\big)\leq\psi(E(x)), \tag{1}$$

where, for some positive constant $k=k(n)$,

$$\psi(t)=\begin{cases} k^2\exp(4\pi t), & n=2, \quad -\infty<t<\infty, \\ k^2t^{2(n-1)/(n-2)}, & n>2, \quad 0<t<\infty, \end{cases}$$

and that $\{x\in\mathbf{R}^2(a)\colon E(x)=-1/\varepsilon\}$ and $\{x\in\mathbf{R}^n(a)\colon E(x)=\varepsilon\}$ $(n>2)$ are compact for each small $\varepsilon>0$ (cf. Levine and Payne [10]). We note that if $n>2$, $\Omega=B(x_0,1)$ (the unit ball) and $E(x)$ is the Green function associated with P in $B(x_0,1)$, then inequality (1) holds for $|x-x_0|\leq r<1$ (see Aviles [3, Lemma 2.2]). We introduce the smooth function $\rho(x)$ $(x\in\mathbf{R}^n(a))$ defined by

$$\rho(x)=\begin{cases} \exp(-2\pi E(x)), & n=2, \\ (\sigma_n(n-2)E(x))^{1/(2-n)}, & n>2, \end{cases}$$

and define a "P-sphere" as a set $S_r\equiv\{x\in\mathbf{R}^n(a)\colon\rho(x)=r\}$. We note that S_r is compact for large r. Since $\partial\mathbf{R}^n(a+\varepsilon_1)$ $(\varepsilon_1>0)$ is compact, $E(x)$ is bounded on $\partial\mathbf{R}^2(a+\varepsilon_1)$ for $n=2$ and there exists a number $\varepsilon_2>0$ such that $E(x)>\varepsilon_2$ on $\partial\mathbf{R}^n(a+\varepsilon_1)$ for $n>2$. Hence, there exists a number $r_0>0$ such that $\partial\mathbf{R}^n(a+\varepsilon_1)\subset\{x\in\mathbf{R}^n(a)\colon\rho(x)<r_0\}$, and consequently we have $\{x\in\mathbf{R}^n(a)\colon\rho(x)>r_0\}\subset\mathbf{R}^n(a)$.

Let $G(t)$ be a real-valued function of class $C^2(\mathbf{R}^1)$. For each function $u\in\mathfrak{N}_L(\Omega)$ we define the function $M[u](r)$ by

$$M[u](r)\equiv\frac{1}{\sigma_n r^{n-1}}\int_{S_r}G(u)\frac{(\nabla\rho(x))^TA(x)(\nabla\rho(x))}{|\nabla\rho(x)|}d\sigma, \qquad r>r_0,$$

where $\nabla\rho(x)$ denotes the gradient of $\rho(x)$, σ denotes the measure on S_r and σ_n denotes the surface area of the unit sphere in $\mathbf{R}^n$, i.e. $\sigma_n=2\pi^{n/2}/\Gamma(n/2)$.

The following fundamental lemma is due to Suleĭmanov [17]. Since the proof is omitted in [17], we give it here.

LEMMA 2.1 (Suleĭmanov [17]). *If $u\in\mathfrak{N}_L(\Omega)$, then we obtain*

$$\sigma_n\frac{d}{dr}\left(r^{n-1}\frac{d}{dr}(M[u](r))\right) \tag{2}$$
$$=\int_{S_r}\left[G''(u)(\nabla u)^TA(x)(\nabla u)+G'(u)P[u]\right]|\nabla\rho|^{-1}d\sigma, \qquad r>r_0.$$

Proof. Since $\nabla\rho=-\sigma_n\rho^{n-1}\nabla E$, we have

$$M[u](r)=-\int_{S_r}G(u)\frac{(\nabla E)^TA(x)(\nabla\rho)}{|\nabla\rho|}d\sigma. \tag{3}$$

From Green's theorem it follows that

(4)

$$M[u](r) = -\int_{r_0<\rho(x)<r} G'(u)(\nabla u)^T A(x)(\nabla E)\,dx - \int_{S_{r_0}} G(u)\frac{(\nabla E)^T A(x)(\nabla \rho)}{|\nabla \rho|}\,d\sigma_0$$

$$= -\int_{r_0}^{r} d\eta \left(\int_{\rho(x)=\eta} G'(u)(\nabla u)^T A(x)(\nabla E)|\nabla \rho|^{-1} d\sigma \right)$$

$$- \int_{S_{r_0}} G(u)\frac{(\nabla E)^T A(x)(\nabla \rho)}{|\nabla \rho|}\,d\sigma_0.$$

Differentiating (4), we obtain

(5)
$$\frac{d}{dr}M[u](r) = -\int_{S_r} G'(u)(\nabla u)^T A(x)(\nabla E)|\nabla \rho|^{-1} d\sigma$$

$$= (\sigma_n r^{n-1})^{-1} \int_{S_r} G'(u)(\nabla u)^T A(x)(-\sigma_n \rho^{n-1}\nabla E)|\nabla \rho|^{-1} d\sigma$$

$$= (\sigma_n r^{n-1})^{-1} \int_{S_r} G'(u)(\nabla u)^T A(x)\frac{\nabla \rho}{|\nabla \rho|}\,d\sigma.$$

A simple computation yields

(6)
$$\int_{r_0<\rho(x)<r} \operatorname{div}\big(G'(u)(\nabla u)^T A(x)\big)\,dx$$

$$= \int_{r_0<\rho(x)<r} \big[G''(u)(\nabla u)^T A(x)(\nabla u) + G'(u)P[u]\big]\,dx.$$

On the other hand, by the divergence theorem and (5), we get

(7)
$$\int_{r_0<\rho(x)<r} \operatorname{div}\big(G'(u)(\nabla u)^T A(x)\big)\,dx$$

$$= \int_{S_r} G'(u)(\nabla u)^T A(x)\frac{\nabla \rho}{|\nabla \rho|}\,d\sigma - \int_{S_{r_0}} G'(u)(\nabla u)^T A(x)\frac{\nabla \rho}{|\nabla \rho|}\,d\sigma_0$$

$$= \sigma_n r^{n-1}\frac{d}{dr}(M[u](r)) - \sigma_n r_0^{n-1}\frac{d}{dr}(M[u](r))\Big|_{r=r_0}.$$

Combining (6) with (7) yields

(8)
$$\sigma_n r^{n-1}\frac{d}{dr}(M[u](r)) - \sigma_n r_0^{n-1}\frac{d}{dr}(M[u](r))\Big|_{r=r_0}$$

$$= \int_{r_0<\rho(x)<r} \big[G''(u)(\nabla u)^T A(x)(\nabla u) + G'(u)P[u]\big]\,dx.$$

Differentiating both sides of (8), we obtain the desired identity (2).

Remark 1. In the case where $P=\Delta$, we choose

$$E(x) = \begin{cases} (2\pi)^{-1}\log(|x|^{-1}), & n=2, \\ (\sigma_n(n-2))^{-1}|x|^{2-n}, & n>2. \end{cases}$$

In this case we have $\rho(x)=|x|$ and $|\nabla\rho|=1$. Hence, $M[u](r)$ reduces to the spherical mean of $G(u)$ over $\{x\in\mathbf{R}^n: |x|=r\}$. We note that Lemma 2.1 with $P=\Delta$ and $G(t)\equiv t$ was established by Noussair and Swanson [12].

LEMMA 2.2. *Defining*

$$K(r)\equiv\int_{S_r}(\nabla\rho)^T A(x)(\nabla\rho)\big(\sigma_n r^{n-1}|\nabla\rho|\big)^{-1}d\sigma,$$

we conclude that $K(r)$ is a positive constant independent of r. Furthermore, if $P=\Delta$ or $\Omega=\mathbf{R}^n$, then we get $K(r)=1$.

Proof. Since $\nabla\rho\not\equiv 0$, we find that $K(r)>0$. Hence, it is sufficient to prove that $K(r)$ is a constant independent of r. An easy computation shows that

$$K(r)=\int_{S_r}\big(-\sigma_n\rho^{n-1}\nabla E\big)^T A(x)(\nabla\rho)\big(\sigma_n r^{n-1}|\nabla\rho|\big)^{-1}d\sigma \tag{9}$$

$$=-\int_{S_r}(\nabla E)^T A(x)\frac{\nabla\rho}{|\nabla\rho|}\,d\sigma.$$

Using Green's theorem, we obtain

$$\int_{S_r}(\nabla E)^T A(x)\frac{\nabla\rho}{|\nabla\rho|}\,d\sigma=\int_{\mathbf{R}^n(a)\cap\{\rho(x)<r\}}P[E]\,dx-\int_{\partial\mathbf{R}^n(a)}(\nabla E)^T A(x)\Big(-\frac{x}{a}\Big)\,ds \tag{10}$$

$$=-1+\int_{\partial\mathbf{R}^n(a)}(\nabla E)^T A(x)\Big(\frac{x}{a}\Big)\,ds.$$

Combining (9) with (10) gives

$$K(r)=1-\int_{\partial\mathbf{R}^n(a)}(\nabla E)^T A(x)\Big(\frac{x}{a}\Big)\,ds,$$

where the right-hand side is a constant independent of r. If $\Omega=\mathbf{R}^n$, it is easy to see that

$$\int_{S_r}(\nabla E)^T A(x)\frac{\nabla\rho}{|\nabla\rho|}\,d\sigma=\int_{\rho(x)<r}P[E]\,dx=-1.$$

Hence, in view of (9), we get $K(r)=1$. If $P=\Delta$, we have $\rho(x)=|x|$ and $|\nabla\rho|=1$ as was stated in Remark 1. A direct calculation shows that

$$K(r)=\int_{|x|=r}(\nabla\rho)^T(\nabla\rho)\big(\sigma_n r^{n-1}|\nabla\rho|\big)^{-1}d\sigma$$

$$=\big(\sigma_n r^{n-1}\big)^{-1}\int_{|x|=r}d\sigma=1.$$

Remark 2. If $P=\Delta$, then (1) holds with equality for

$$k=k(n)=\begin{cases}(2\pi)^{-1}, & n=2,\\ \sigma_n^{-1}\big(\sigma_n(n-2)\big)^{(n-1)/(n-2)}, & n>2.\end{cases}$$

3. Superlinear equations. Lemma 2.1 in §2 will be used to derive sufficient conditions for every solution of the superlinear equation (*) to be oscillatory in Ω. By the same arguments as were used by Noussair and Swanson [12], we get the following main theorem:

THEOREM 3.1. *Assume that the following conditions are satisfied*:

(i) $c(x,-\xi)=-c(x,\xi)$ *for all* $(x,\xi)\in\Omega\times(0,\infty)$.

(ii) $c(x,\xi)\geq q(\rho(x))\phi(\xi)$ *for all* $(x,\xi)\in\Omega\times(0,\infty)$, *where* q *is continuous and positive in* $[r_0,\infty)$ *and* ϕ *is continuous, positive and convex in* $(0,\infty)$.

Then every solution u *of* $(*)$ *is oscillatory in* Ω *if the ordinary differential inequalities*

$$\frac{d}{dr}\left(r^{n-1}\frac{dz}{dr}\right)+\frac{1}{\beta(n)}r^{n-1}q(r)\phi(z)\leq r^{n-1}F(r), \tag{11}$$

$$\frac{d}{dr}\left(r^{n-1}\frac{dz}{dr}\right)+\frac{1}{\beta(n)}r^{n-1}q(r)\phi(z)\leq -r^{n-1}F(r), \tag{12}$$

are oscillatory at $r=\infty$ *in the sense that neither* (11) *nor* (12) *has a solution which is positive on* $[r,\infty)$ *for any* $r>r_0$, *where*

$$\beta(n)=\begin{cases}k^2(2\pi)^2, & n=2,\\ k^2\sigma_n^2\left(\sigma_n(n-2)\right)^{2(n-1)/(2-n)}, & n>2\end{cases}$$

and

$$F(r)=\left(K\sigma_n r^{n-1}\right)^{-1}\int_{S_r}f(x)|\nabla\rho|^{-1}d\sigma,$$

where K *is the positive constant defined in Lemma* 2.2.

Proof. Suppose to the contrary that there exists a solution u which has no zero in $\mathbf{R}^n(r_1)$ for some $r_1>0$. First we assume $u>0$ in $\mathbf{R}^n(r_1)$. For the number r_1 there exists a number $r_2>0$ such that $\{x\in\mathbf{R}^n:\rho(x)>r_2\}\subset\mathbf{R}^n(r_1)$. Lemma 2.1 with $G(t)\equiv t$ implies that

$$\begin{aligned}\frac{d}{dr}\left(r^{n-1}\frac{dv}{dr}\right)&=(K\sigma_n)^{-1}\int_{S_r}P[u]|\nabla\rho|^{-1}d\sigma\\ &=(K\sigma_n)^{-1}\int_{S_r}(-c(x,u)+f(x))|\nabla\rho|^{-1}d\sigma,\qquad r>r_2,\end{aligned} \tag{13}$$

where

$$v=v[u](r)=\frac{1}{K\sigma_n r^{n-1}}\int_{S_r}u\,\frac{(\nabla\rho)^TA(x)(\nabla\rho)}{|\nabla\rho|}d\sigma.$$

By hypothesis (ii) of Theorem 3.1 we get

$$\begin{aligned}(K\sigma_n)^{-1}\int_{S_r}c(x,u)\,d\sigma&\geq(K\sigma_n)^{-1}q(r)\int_{S_r}\phi(u)|\nabla\rho|^{-1}d\sigma\\ &=r^{n-1}q(r)\int_{S_r}\phi(u)\left(K\sigma_n r^{n-1}|\nabla\rho|\right)^{-1}d\sigma.\end{aligned} \tag{14}$$

Using inequality (1), we easily obtain

$$(\nabla\rho)^TA(x)(\nabla\rho)\leq\beta(n).$$

Hence, it is true that

(15)

$$\int_{S_r}\phi(u)\left(K\sigma_n r^{n-1}|\nabla\rho|\right)^{-1}d\sigma\geq\frac{1}{\beta(n)}\int_{S_r}\phi(u)(\nabla\rho)^TA(x)(\nabla\rho)\left(K\sigma_n r^{n-1}|\nabla\rho|\right)^{-1}d\sigma.$$

Application of Jensen's inequality [14, p. 160] gives

$$\int_{S_r} \phi(u)(\nabla\rho)^T A(x)(\nabla\rho)\left(K\sigma_n r^{n-1}|\nabla\rho|\right)^{-1} d\sigma \geq \phi(v[u](r)). \tag{16}$$

Combining (14)–(16), we have

$$(K\sigma_n)^{-1}\int_{S_r} c(x,u)\,d\sigma \geq r^{n-1}q(r)\,\frac{1}{\beta(n)}\,\phi(v[u](r)). \tag{17}$$

From (13) and (17) it follows that

$$\frac{d}{dr}\left(r^{n-1}\frac{dv}{dr}\right) \leq -\frac{1}{\beta(n)}r^{n-1}q(r)\phi(v)+r^{n-1}F(r),$$

which is equivalent to (11). Hence, $v[u](r)$ is a positive solution of (11) in (r_2,∞). If $u<0$ in $\mathbf{R}^n(r_1)$, $U\equiv -u$ is a positive solution of the equation $L[U]=-f(x)$. Hence, $v[U](r)$ is a positive solution of (12) in (r_2,∞). This contradicts the hypothesis and completes the proof.

Using a result of Kusano and Naito [9, Thms. 2 and 3], we obtain the following results.

THEOREM 3.2. *Assume that $c(x,\xi)$ satisfies hypotheses* (i) *and* (ii) *of Theorem* 3.1. *Moreover, assume that if $n=2$, then for all large $\tilde{r}$*

$$\liminf_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\frac{\log s}{\log r}\right)sF(s)\,ds=-\infty,$$

$$\limsup_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\frac{\log s}{\log r}\right)sF(s)\,ds=\infty,$$

and if $n>2$, then for all large $\tilde{r}$

$$\liminf_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\left(\frac{s}{r}\right)^{n-2}\right)sF(s)\,ds=-\infty,$$

$$\limsup_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\left(\frac{s}{r}\right)^{n-2}\right)sF(s)\,ds=\infty.$$

Then every solution u of $(*)$ is oscillatory in Ω.

THEOREM 3.3. *Assume that $c(x,\xi)$ satisfies hypotheses* (i) *and* (ii) *of Theorem* 3.1. *Moreover, assume that the ordinary differential inequality*

$$\frac{d}{dr}\left(r^{3-n}\frac{d}{dr}(r^{n-2}z)\right)+\frac{1}{\beta(n)}rq(r)\phi(z)\leq 0 \tag{18}$$

has no eventually positive solution, and that there exists a C^2-function $\theta:[r_0,\infty)\to\mathbf{R}^1$ with the following properties:

(i) *$\theta(r)$ takes both positive and negative values for arbitrarily large r.*

(ii) $\frac{d}{dr}(r^{3-n}\frac{d}{dr}(r^{n-2}\theta(r)))=rF(r)$, $r\geq r_0$.

(iii) $\lim_{r\to\infty} r^{n-2}\theta(r)=0$.

Then every solution u of $(*)$ is oscillatory in Ω.

Example 1. We consider the uniformly elliptic equation

$$\sum_{i,j=1}^{3} D_i\left(a_{ij}(x)D_j u\right)+u^5=(\nabla\rho)^T A(x)(\nabla\rho)(\log\rho)\sin\rho, \qquad x\in\Omega. \tag{19}$$

Here $n=3$ and $q(\rho(x))\equiv 1$. It is easily seen from Lemma 2.2 that $F(r)=(\log r)\sin r$. Since

$$\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(\log s)\sin s\,ds=-(\log r)\sin r+B(r,\tilde{r}),$$

where $B(r,\tilde{r})$ is bounded as $r\to\infty$, we have

$$\liminf_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(\log s)\sin s\,ds=-\infty,$$

$$\limsup_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(\log s)\sin s\,ds=\infty.$$

Hence, it follows from Theorem 3.2 that every solution u of (19) is oscillatory in Ω.

Example 2. We consider the uniformly elliptic equation

$$\sum_{i,j=1}^{3}D_i\left(a_{ij}(x)D_ju\right)+\rho^4u^7=2(\nabla\rho)^TA(x)(\nabla\rho)e^{-\rho}\sin\rho,\qquad x\in\Omega. \tag{20}$$

Here $n=3$ and $q(r)=r^4$. Since

$$\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(2e^{-s}\sin s)\,ds<\infty\qquad(r\to\infty),$$

Theorem 3.2 does not apply to (20), but Theorem 3.3 does. In this case the ordinary differential inequality (18) becomes

$$\frac{d^2}{dr^2}(rz)+\frac{1}{\beta(3)}r^5z^7\le 0. \tag{21}$$

The above inequality may be reduced by the substitution $w=rz$ to

$$\frac{d^2w}{dr^2}+\frac{1}{\beta(3)}r^{-2}w^7\le 0. \tag{22}$$

Since

$$\int^{\infty}r\left(\frac{1}{\beta(3)}r^{-2}\right)dr=\infty,$$

the ordinary differential equation

$$\frac{d^2w}{dr^2}+\frac{1}{\beta(3)}r^{-2}w^7=0$$

has no eventually positive solution (see [6], [11]). Using results of [5], [15], we find that inequality (22) has no eventually positive solution, and therefore inequality (21) has no eventually positive solution. It is easy to see that $\theta(r)\equiv r^{-1}e^{-r}(-\sin r+\cos r+r\cos r)$ satisfies hypotheses (i)–(iii) of Theorem 3.3. Hence, every solution u of (20) is oscillatory in Ω.

Remark 3. In the case where $P=\Delta$, it is easily seen from Remarks 1 and 2 in §2 that $\beta(n)=1$ and $(\nabla\rho)^TA(x)(\nabla\rho)=1$.

4. Sublinear equations. The sublinear case was first discussed by Kitamura and Kusano [7], and recently by Kura [8], Noussair and Swanson [13] and Onose [16]. We consider the case where $c(x,u)=c(x)\Phi(u)$ and $f(x)\equiv 0$, i.e.,

$$L[u]\equiv\sum_{i,j=1}^{n}D_i\left(a_{ij}(x)D_ju\right)+c(x)\Phi(u)=0,\qquad x\in\Omega. \tag{**}$$

We assume that the following conditions are satisfied in this section:

(B-I) $\Phi(\xi)$ is a real-valued function of class $C(\mathbf{R}^1)\cap C^1(\mathbf{R}^1\setminus\{0\})$;

(B-II) $\xi\Phi(\xi)>0$ and $\Phi'(\xi)\geq 0$ for $\xi\neq 0$;

(B-III) $\int_0^{\pm\varepsilon} d\xi/\Phi(\xi)<\infty$ for some $\varepsilon>0$.

The n-dimensional sublinear Emden–Fowler equation

$$\sum_{i,j=1}^{n} D_i\big(a_{ij}(x)D_j u\big)+c(x)|u|^{\gamma}\operatorname{sgn} u=0, \qquad 0<\gamma<1,$$

is an important special case of (**) which satisfies (B-I)–(B-III).

THEOREM 4.1. *Under assumptions* (B-I)–(B-III), *every solution u of* (**) *is oscillatory in Ω if*

$$\limsup_{r\to\infty}\int_{\tilde r}^{r}\left(1-\frac{\theta_n(s)}{\theta_n(r)}\right)s\tilde c(s)\,ds=\infty$$

for some $\tilde r>r_0$, where

$$\theta_n(r)=\begin{cases}\log r, & n=2,\\ r^{n-2}, & n>2,\end{cases}$$

and

$$\tilde c(r)=\big(K\sigma_n r^{n-1}\big)^{-1}\int_{S_r} c(x)|\nabla\rho|^{-1}d\sigma.$$

Proof. Suppose to the contrary that there exists a solution u of (**) which is either positive or negative in $\{x\in\mathbf{R}^n:\rho(x)\geq r_2\}$ for some $r_2>0$. Defining

$$G(u)\equiv\int_0^u\frac{d\xi}{\Phi(\xi)},$$

we see that $G(u)>0$, $G'(u)=\Phi(u)^{-1}$ and $G''(u)=-\Phi'(u)/\Phi(u)^2\geq 0$. Hence, Lemma 2.1 in §2 implies that

$$\begin{aligned}\frac{d}{dr}\left(r^{n-1}\frac{d}{dr}(M[u](r))\right)&\leq\sigma_n^{-1}\int_{S_r}\Phi(u)^{-1}P[u]|\nabla\rho|^{-1}d\sigma\\ &=-\sigma_n^{-1}\int_{S_r}c(x)|\nabla\rho|^{-1}d\sigma=-Kr^{n-1}\tilde c(r), \qquad r\geq r_2,\end{aligned}$$

or equivalently

$$\frac{d}{dr}\left(r^{3-n}\frac{d}{dr}\big(r^{n-2}M[u](r)\big)\right)\leq -Kr\tilde c(r), \qquad r\geq r_2.$$

Using the same arguments as in [8, Thm. 1], we conclude that

$$0\leq\liminf_{r\to\infty}\frac{r^{n-2}M[u](r)}{\theta_n(r)}\leq d_n-c_n\limsup_{r\to\infty}\int_{r_2}^{r}\left(1-\frac{\theta_n(s)}{\theta_n(r)}\right)Ks\tilde c(s)\,ds=-\infty,$$

where $c_n=1$ $(n=2)$, $c_n=(n-2)^{-1}$ $(n>2)$ and $d_n=c_n r^{3-n}\frac{d}{dr}(r^{n-2}M[u](r))|_{r=r_2}$. This contradiction establishes the theorem.

COROLLARY. *Under assumptions* (B-I)–(B-III), *every solution u of* (**) *is oscillatory in Ω if*

$$\int_{\tilde r}^{\infty}s\tilde c(s)\,ds=\infty \quad \textit{for some } \tilde r>r_0.$$

Proof. Since

$$\int_{\tilde{r}}^{r}\left(1-\frac{\theta_n(s)}{\theta_n(r)}\right)s\tilde{c}(s)\,ds=\frac{1}{\theta_n(r)}\int_{\tilde{r}}^{r}\left(\int_{s}^{r}\zeta(t)\,dt\right)s\tilde{c}(s)\,ds$$

$$=\frac{1}{\theta_n(r)}\int_{\tilde{r}}^{r}\zeta(t)\,dt\int_{\tilde{r}}^{t}s\tilde{c}(s)\,ds\to\infty\qquad(r\to\infty),$$

where $\zeta(t)=t^{-1}(n=2)$ and $\zeta(t)=(n-2)t^{n-3}$ $(n>2)$, the conclusion follows from Theorem 4.1.

Example 3. We consider the sublinear equation

$$(23)\qquad \sum_{i,j=1}^{3} D_i\big(a_{ij}(x)D_j u\big)+(\nabla\rho)^T A(x)(\nabla\rho)\rho(\sin\rho)|u|^{\gamma}\operatorname{sgn}u=0,\qquad 0<\gamma<1.$$

In this case we get $\tilde{c}(r)=r\sin r$. Since

$$\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(s\sin s)\,ds=-r\sin r+B(r,\tilde{r}),$$

where $B(r,\tilde{r})$ is bounded as $r\to\infty$, it follows that

$$\limsup_{r\to\infty}\int_{\tilde{r}}^{r}\left(1-\frac{s}{r}\right)s(s\sin s)\,ds=\infty.$$

Hence, Theorem 4.1 implies that every solution u of (23) is oscillatory in Ω.

Acknowledgment. The author is grateful to the referee who has kindly let him know about the work of Aviles [3].

REFERENCES

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

[2] _____, *Oscillation criteria for semilinear equations in general domains*, Canad. Math. Bull., 19 (1976), pp. 137–144.

[3] P. AVILES, *A study of the singularities of solutions of a class of nonlinear elliptic partial differential equations*, Comm. Partial Differential Equations, to appear.

[4] S. ITÔ, *Fundamental solutions of parabolic differential equations and boundary value problems*, Japan. J. Math., 27 (1957), pp. 55–102.

[5] A. G. KARTSATOS, *On nth order differential inequalities*, J. Math. Anal. Appl., 52 (1975), pp. 1–9.

[6] I. T. KIGURADZE, *Oscillation properties of solutions of certain ordinary differential equations*, Dokl. Akad. Nauk SSSR, 144 (1962), pp. 33–36, Soviet Math. Dokl., 3 (1962), pp. 649–652.

[7] Y. KITAMURA AND T. KUSANO, *An oscillation theorem for a sublinear Schrödinger equation*, Utilitas Math., 14 (1978), pp. 171–175.

[8] T. KURA, *Oscillation criteria for a class of sublinear elliptic equations of the second order*, Utilitas Math., to appear.

[9] T. KUSANO AND M. NAITO, *Oscillation criteria for a class of perturbed Schrödinger equations*, Canad. Math. Bull., 25 (1982), pp. 71–77.

[10] H. A. LEVINE AND L. E. PAYNE, *On the nonexistence of entire solutions to nonlinear second order elliptic equations*, this Journal, 7 (1976), pp. 337–343.

[11] I. LIČKO AND M. ŠVEC, *Le caractère oscillatoire des solutions de l'équation* $y^{(n)}+f(x)y^{\alpha}=0$, $n>1$, Czechoslovak Math. J., 13 (1963), pp. 481–491.

[12] E. S. NOUSSAIR AND C. A. SWANSON, *Oscillation theory for semilinear Schrödinger equations and inequalities*, Proc. Roy. Soc. Edinburgh Sect. A, 75 (1976), pp. 67–81.

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

[14] G. O. OKIKIOLU, *Aspects of the Theory of Bounded Integral Operators in L^p-spaces*, Academic Press, New York, 1971.

[15] H. ONOSE, *A comparison theorem and the forced oscillation*, Bull. Austral. Math. Soc., 13 (1975), pp. 13–19.

[16] _____, *Oscillation criteria for the sublinear Schrödinger equation*, Proc. Amer. Math. Soc., 85 (1982), pp. 69–72.

[17] N. M. SULEĬMANOV, *On the behavior of solutions of second order nonlinear elliptic equations with linear principal part*, Dokl. Akad. Nauk SSSR, 247 (1979), pp. 805–809, Soviet Math. Dokl., 20 (1979), pp. 815–819.

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

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0007 $01.25/0

ASYMPTOTIC SOLUTIONS FOR A DIRICHLET PROBLEM WITH AN EXPONENTIAL NONLINEARITY*

J. L. MOSELEY†

Abstract. We consider the two-dimensional nonlinear Dirichlet problem

$$-\Delta u=\lambda e^{u}, \qquad y\in\Omega,$$
$$u=\phi, \qquad y\in\partial\Omega,$$

where $y=(y_1,y_2)$, Δ is the Laplacian operator, Ω is a simply connected region bounded by a smooth closed Jordan curve, the boundary data ϕ is continuous and λ is positive. Our primary concern is with obtaining the large norm (second) solution for λ tending to 0_+. This is accomplished by obtaining an asymptotic solution which is used as a first approximation for a modified Newton's method. In this paper we examine the implicit constraints previously required for $\phi\equiv 0$ and extend the results to the case of nonzero boundary data.

AMS-MOS subject classification (1980). Primary 35J25, 35J60, 35J65

Key words. elliptic equation, nonlinear boundary value problem

Introduction. We consider the two-dimensional nonlinear Dirichlet problem

(P)
$$-\Delta u(y)=\lambda e^{u(y)}, \qquad y\in\Omega,$$
$$u(y)=\phi(y), \qquad y\in\partial\Omega,$$

where $y=(y_1,y_2)$, Δ is the Laplacian operator, Ω is a simply connected region bounded by a smooth closed Jordan curve, the boundary data $\phi(y)$ is continuous and λ is positive. Our concern is with obtaining classical solutions ($u\in C^2(\Omega)\cap C(\bar{\Omega})$) for λ tending to 0_+, particularly the large norm (second) solution [17]. As in [17], this is accomplished by obtaining an asymptotic solution which is used as a first approximation for a modified Newton's method. The purpose of this paper is to examine closely the implicit constraints required in [17] for the case $\phi\equiv 0$ and to extend the results to the case of nonzero boundary data.

Problems of type (P) have application in a wide variety of areas including the theory of thermal ignition of a chemically active mixture of gases [6], the theory of nonlinear diffusion by nonlinear sources [10] and the study of Riemann surfaces with bounded Gaussian curvature [9].

Section 1 summarizes the results of this paper as well as applicable results from the literature. Sections 2 through 9 give the requisite development for the proof of Theorem 1 of §1.

1. Summary of results. The outstanding feature of (P) is that $-\Delta u=\lambda e^{u}$ is analytic so that all solutions are analytic in Ω [13]. If, in addition, the boundary $\partial\Omega$ and the boundary data ϕ are analytic, all solutions to the boundary value problem are analytic in $\bar{\Omega}$ [13]; however, this is too restrictive for our purposes.

To obtain an equivalent problem with zero boundary data, we let $u_1=u-u_0$, where u_0 is the solution of the associated harmonic problem ($\lambda=0$):

(P′)
$$-\Delta u_1=\lambda e^{u_0+u_1}, \qquad y\in\Omega,$$
$$u_1=0, \qquad y\in\partial\Omega,$$

*Received by the editors October 1, 1980, and in final revised form July 30, 1982. This work was supported in part by a grant from the West Virginia University Senate.

†Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506.

with $\lambda>0$. The existence of a solution of (P′) for λ sufficiently small is well known ([4, p. 373]). All solutions of (P′) are superharmonic and hence positive.

The technique of Keller and Cohen [10] can be applied to (P′) for any λ in the "spectrum" (the set of all $\lambda>0$ such that (P′) has a solution) to obtain the minimal solution. Keller and Cohen's results also show that the spectrum is a finite interval with least upper bound λ^* satisfying $\lambda^*\leq\mu_1$, where μ_1 is the least eigenvalue of the linear problem

$$\begin{aligned} -\Delta v&=\mu e^{u_0}v, && y\in\Omega,\\ v&=0, && y\in\partial\Omega. \end{aligned}$$

Bandle's results [2] show that for some $\varepsilon>0$

$$\lambda^*\geq\frac{4\pi}{A}e^{-1-\hat{u}_0}+\varepsilon,$$

where $\hat{u}_0=\max_{y\in\Omega}(u_0(y))$ and A is the area of Ω. For $\phi\equiv 0$ she has also shown that [3]

$$\frac{2\pi}{A}\leq\lambda^*\leq\frac{2}{R_0^2},$$

where R_0 is the maximal conformal radius of Ω. For this case she gives the estimates

$$1-\sqrt{1-\frac{\lambda}{2}R_y^2}\leq 2e^{-u(y)/2}\leq 1+\sqrt{1-\frac{\lambda}{2}R_y^2},$$

where R_y is the conformal radius of y with respect to Ω.

Crandall and Rabinowitz's results [5] show that λ^* is in the spectrum and that there are two solutions for every $\lambda\in(0,\lambda^*)$. This second result is obtained by applying a nonconstructive topological technique having its origin in the Ljusternik–Schnirelman theory of critical points.

For the case $\phi\equiv 0$ and $\partial\Omega$ a circle, we have that the symmetric solutions of (P′) are given by

$$e^{u_0}=\frac{b_0}{\left(1+b_0\lambda r^2/8\right)^2} \quad \text{with } b_0=\frac{32}{\lambda^2R^4}\left(1-\frac{\lambda R^2}{4}-\sqrt{1-\frac{\lambda R^2}{2}}\right),$$

$$e^{u_1}=\frac{b_1}{\left(1+b_1\lambda r^2/8\right)^2} \quad \text{with } b_1=\frac{32}{\lambda^2R^4}\left(1-\frac{\lambda R^2}{4}+\sqrt{1-\frac{\lambda R^2}{2}}\right),$$

where R is the radius of the circle and r is the radial variable [2]. In this case $\lambda^*=2/R^2$ and it is easy to see that u_0 is the minimal solution, u_1 increases without bound at the origin as $\lambda\to 0$ and the two solutions coincide at λ^*. By a technique described in [7] it can be shown that these are the only solutions.

The general integral of $-\Delta u=\lambda e^u$ may be written as

$$\lambda e^u=\frac{8|F'(w)|^2}{\left(1+|F(w)|^2\right)^2},$$

where F is a meromorphic function of $w=y_1+iy_2$ whose Schwarzian derivative is holomorphic (see [11] and [14, Appendix A]). Weston [17] used this to develop an asymptotic approximation of the "large norm" solution when Ω is an arbitrary simply connected bounded domain with smooth boundary (continuously turning tangent) and $\phi\equiv 0$. However, three additional implicit constraints on the domain (in terms of the

conformal map of the unit disc U onto Ω) are required. As with u_1 given above, the distinctive feature of the asymptotic solution is the existence of a single maximum at which the solution increases without bound as $\lambda \to 0$. For λ sufficiently small, Weston also showed that if the asymptotic solution is used as a first approximation in an appropriate modified Newton's iteration scheme, then an exact large norm solution is generated provided that the asymptotic solution is taken to order greater than or equal to three.

In this paper we examine more closely the three implicit constraints given in [17] and extend the results to the case of nonzero ϕ. Specifically, we remove one of the constraints by using a different form of the general integral. The two remaining constraints are associated with the order of the asymptotic solution.

As given by Weston, the first order constraint requires the existence of a complex number δ in the unit disc $U=\{z\in\mathbb{C}: |z|<1\}$ such that

$$(1-|\delta|^2)f''(\delta)-2\bar{\delta}f'(\delta)=0, \tag{1.1}$$

where f is a conformal map of U onto Ω with $\overline{\Omega}$ the homeomorphic image of $\overline{U}$ (i.e., f characterizes $\overline{\Omega}$).

As no difficulty will arise, we will consider a region $\Omega\subseteq\mathbb{R}^2$ to also be a subset of $\mathbb{C}$ throughout this paper, using $y=(y_1,y_2)\in\Omega\subset\mathbb{R}^2$ and $w=y_1+iy_2\in\Omega\subseteq\mathbb{C}$. When the region is the unit disc U, we will use $x=(x_1,x_2)\in U\subseteq\mathbb{R}^2$ and $z=x_1+ix_2\in\mathbb{C}$. For any region $\Omega\subseteq\mathbb{C}$ with smooth boundary, we let $H(\Omega)$ be the holomorphic functions on Ω, $A(\Omega)$ be the functions in $H(\Omega)$ which are continuous on $\overline{\Omega}$, $A_\alpha(\Omega)$ be the functions in $A(\Omega)$ that satisfy a Lipschitz condition (as a function of arclength) of order α on $\partial\Omega$, $0<\alpha\leq 1$, and $A_N(\Omega)$ be the functions in $A(\Omega)$ which are nonzero in $\overline{\Omega}$. Additionally, H^p, $0<p\leq\infty$, are the usual Hardy spaces on U and we let H_N be the functions in H^∞ which are nonzero in U. If the region is not specified, it is assumed to be U.

If f is the conformal map of U onto Ω, then $f'\in H^p$ for $0<p<\infty$ [8, p. 425].

LEMMA. *If $f'\in H_N$, then $\delta_0\in U$ is a solution of* (1.1) *if and only if*

$$R(\delta)=|f'(\delta)|(1-|\delta|^2)$$

has a relative extremum at δ_0.

If f is the conformal map of U onto Ω, then $R(\delta)$ is the conformal radius of Ω at $f(\delta)$. Thus, $\delta_0=f^{-1}(w_0)$ satisfies (1.1), where w_0 is a point of maximal conformal radius for Ω. If δ_0 is any solution of (1.1), then the use of a normalized conformal map f_N such that $f_N(0)=f(\delta_0)$ converts (1.1) to

$$f_N''(0)=0. \tag{1.2}$$

This simplifies the computation of the asymptotic solution as well as the modified Newton's method. Some examples where (1.1) can be solved explicitly are given in §7.

Using f_N, the second order constraint in [17] simplifies to

$$|f_N'''(0)|\neq 2|f_N'(0)|. \tag{1.3}$$

If Ω is convex, then an application of the area theorem [15] yields

$$|f_N'''(0)|\leq 2\left(1-\frac{1}{M^2}\right)|f_N'(0)| \tag{1.4}$$

if $M=\sup_{z\in U}|f_N'(z)/f_N'(0)|<\infty$.

To handle the case of arbitrary (continuous) boundary data, we let $h(w)=u_0(y)+iv_0(y)$, where u_0 is the solution of the associated harmonic problem ($\lambda=0$) and v_0 is its conjugate harmonic function. Next we let f_Ω be the conformal map of U onto Ω and

$f_\phi(z)=\exp\{(1/2)h(f_\Omega(z))\}$. The function f_ϕ characterizes the boundary data and we will show that by letting f satisfy $f'=f_\Omega' f_\phi$, we can reduce (P) to the (conformal transplantation of the) zero boundary data case. Again, the first order constraint is the existence of a δ satisfying (1.1).

We refer to f, where $f'=f_\Omega' f_\phi$ as an associated map for (P). Since $f'\in H_N$ implies that (1.1) has a solution δ_0 and $f_\phi\in H_N$ (the real part of H is continuous on $\bar{U}$) we have $f_\Omega'\in H_N$ as a sufficient condition to satisfy the first order constraint for arbitrary boundary data. As before, if δ_0 is any solution of (1.1) for a given associated map a normalized associated map can be obtained such that $f_N(0)=f(\delta_0)$ with (1.2) being satisfied. The second order constraint again becomes (1.3). Sufficient conditions for (1.3) will be given later.

The above can be used to obtain a somewhat simplified asymptotic solution together with an associated modified Newton's method which converges to an exact large norm solution for λ sufficiently small. In order to maintain uniform control over the boundary behavior of our asymptotic solutions (repeated use of Schwarz's formula [1, p. 167] is required) and hence the kernels for the operators used in the modified Newton's method, it is convenient (if not essential) to assume stronger smoothness conditions on ϕ and $\partial\Omega$. As indicated below, we require $f'\in A_N$.

THEOREM 1. *Let the associated map f for* (P) *satisfy $f'\in A_N$ and the normalized associated map f_N satisfy* (1.3). *If g_Ω is the inverse of f_Ω, where $f_N'=f_\Omega' f_\phi$, then an nth order large norm asymptotic solution can be obtained in the form*:

$$e^{-u(w;\lambda)/2}=\frac{|g_\Omega(w)|^2+(\lambda/8)\left|g_\Omega(w)\int^{g_\Omega(w)}[G(\hat{z};\lambda)]^2\left(f_N(\hat{z})/\hat{z}^2\right)d\hat{z}\right|^2}{|G(g_\Omega(w);\lambda)|^2} \tag{1.5}$$

where

$$G(z;\lambda)=1+G_1(z)+\cdots+\lambda^{n-1}G_{n-1}(z),$$

the functions G_i being given in §8. *Furthermore, an exact solution of* (P) *can be generated via a modified Newton's method for a conformal transplantation of* (P) *provided* λ *is sufficiently small and the asymptotic solution is taken to at least order* 3.

We call $\partial\Omega$ Dini-smooth if in addition to being smooth, the tangent angle $\beta(s)$ as a function of arclength s satisfies $|\beta(s_2)-\beta(s_1)|<\omega(s_2-s_1)$ for $s_1<s_2$, where $\omega(x)$ is an increasing function for which $\int_0^1\omega(x)/x\,dx<\infty$. If $\partial\Omega$ is Dini-smooth, then $f_\Omega'(z)\in A_N$ (and $g_\Omega'(w)\in A_N(\Omega)$) [16, p. 298]. Note that if $\beta(s)$ satisfies a Lipschitz condition of order α, $0<\alpha\leq 1$, then $\partial\Omega$ is Dini-smooth and, in fact, $f_\Omega\in A_\alpha$. If, in addition, ϕ satisfies a Lipschitz condition of order α as a function of arclength, then $f_\phi\in A_\alpha$. If $\partial\Omega$ is Dini-smooth and $h=u_0+iv_0$ is in $A(\Omega)$, we say that (P) is conformally smooth. Thus, (P) conformally smooth implies $f_\Omega', f_\phi\in A_N$.

COROLLARY 1. *If $\phi\equiv 0$, Ω is convex and $\partial\Omega$ is Dini-smooth, then the conclusion of Theorem* 1 *is valid.*

COROLLARY 2. *Assume* (P) *is conformally smooth, Ω is convex and f_Ω' and f_ϕ have been normalized separately (thus $f_\Omega''(0)=f_\phi'(0)=0$). If $|f_\phi'(0)/f_\phi(0)|<2/M^2$, where $M=\max_{z\in\bar{U}}|f_\Omega'(z)/f_\Omega'(0)|$, then the conclusion of Theorem* 1 *is valid.*

COROLLARY 3. *Assume* (P) *is conformally smooth, Ω is convex, $\partial\Omega$ is symmetric with respect to two perpendicular axes and f is the odd conformal map of U onto Ω with $f_\Omega(0)=0, f_\Omega'(0)>0$. If ϕ is symmetric with respect to both axes and*

$$|h''(0)|<\frac{4}{M^2}, \tag{1.6}$$

where $M=\max_{z\in\bar{U}}|f_\Omega'(z)|$, *then the conclusion of Theorem* 1 *is valid.*

COROLLARY 4. *The requirement* (1.6) *in Corollary* 3 *may be replaced by* $M_0M^2<4|f'(0)|$, *when* $M_0=\max_{w\in\bar{\Omega}}|h'(w)|<\infty$.

The proofs of the corollaries are straightforward and contained in [14].

2. Solution to the Dirichlet problem via the general integral. The general integral for $-\Delta u=\lambda e^u$ with $\lambda>0$ is given by

$$e^u=\frac{|1/v^2(w)|}{\left(1+(\lambda/8)|\int^w d\hat{w}/v^2(\hat{w})|^2\right)^2}=\frac{|F'(w)|^2}{\left(1+(\lambda/8)|F(w)|^2\right)^2} \tag{2.1}$$

or

$$e^{-u/2}=|v(w)|^2+\frac{\lambda}{8}\left|v(w)\int^w\frac{dw}{v^2(w)}\right|^2 \tag{2.2}$$

where, as a function of $w=y_1+iy_2$, v is holomorphic, F is meromorphic and $F'(w)=1/v^2(w)\neq 0$ (see [11] and [14, Appendix A]).

This representation holds globally in a simply connected region Ω provided the equivalent conditions:

i)
$$S(w)=\left(\frac{F''(w)}{F'(w)}\right)'-\frac{1}{2}\left(\frac{F''(w)}{F'(w)}\right)^2=-\frac{2v''(w)}{v(w)}$$
is holomorphic in Ω, where $S(w)=\{F(w);w\}$ is the Schwarzian derivative of F;

ii) $v(w)$ has at most simple zeros in Ω and if $v(w_0)=0$, then $v''(w_0)=0$;

iii) $F(w)$ is holomorphic in Ω except for simple poles

are satisfied. Note that u has removable singularities at the zeros of v (poles of F).

The use of (2.1) or (2.2) rather than the general integral given in §1 removes one of the constraints in [17]. Also, the computations required to obtain u asymptotically make (2.2) more desirable than (2.1). We, therefore, formulate our results in terms of v using (2.2) and require ii).

For $\lambda>0$ let $B_\lambda(\Omega)$ be the set of all $v(w)=v(w;\lambda)$ in $H(\Omega)$ such that $u(w)=u(w;\lambda)$ as defined by (2.2) is continuous in $\bar{\Omega}$; thus $v\in B_\lambda(\Omega)$ implies ii). Note that $A(\Omega)\subseteq B_\lambda(\Omega)$ provided ii) is also satisfied.

A classical solution of (P) is obtained provided $v(w)\in B_\lambda(w)$ can be found to satisfy the boundary condition:

$$e^{-\phi(w)/2}=|v(w)|^2+\frac{\lambda}{8}\left|v(w)\int^w\frac{d\hat{w}}{(v(\hat{w}))^2}\right|^2,\qquad w\in\partial\Omega. \tag{C1}$$

When (P) is solvable, the existence of some v is assured, but in general, v will not be unique (for example, see [14, Appendix B]). This will not be a problem (but rather a blessing) as our concern is with finding v's which yield solutions of (P), in particular, the large norm solution for $\lambda\to 0_+$.

To illustrate this approach we first consider the harmonic case ($\lambda=0$):

$$\Delta u=0,\quad w\in\Omega,\qquad u=\phi,\quad w\in\partial\Omega. \tag{L1}$$

To solve (L1) we seek $v\in B_0(\Omega)=\{v\in H(\Omega):\ln|v|$ is continuous in $\bar{\Omega}\}$ satisfying the boundary condition $\phi(w)=-2\ln|v(w)|^2$ for $w\in\partial\Omega$.

For the case where Ω is the unit disk $U=\{z\in\mathbb{C}:|z|<1\}$ (ϕ continuous), the solution of (L1) is given by $u(z)=\mathrm{Re}(h(z))$, where

$$h(z)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{e^{i\theta}+z}{e^{i\theta}-z}\right)\phi(e^{i\theta})\,d\theta+ic \tag{2.3}$$

$$=\frac{1}{\pi i}\oint_{\partial u}\left(\frac{1}{\hat{z}-z}-\frac{1}{2\hat{z}}\right)\phi(\hat{z})\,d\hat{z}+ic$$

and the imaginary part of the integral is interpreted in the Cauchy principal-value sense if necessary. Defining $f_\phi=e^{(1/2)h(z)}$ we obtain $\phi(z)=\mathrm{Re}\,h(z)=\ln|f_\phi(z)|^2$ for $z\in\partial U$. Hence, we may take $v(z)=(f_\phi(z))^{-2}=e^{-h(z)}$, $z\in U$. Note that the continuity of ϕ will not assure that $v\in A_N$ (e.g., the imaginary part of h may diverge to $+\infty$ [18, p. 253]). However, $h\in A$ implies $v\in A_N$.

For arbitrary Ω we consider a conformal transplantation to the unit disc. As before, let $f_\Omega\in A$ map $\overline{U}$ onto $\overline{\Omega}$ and $g_\Omega\in A(\Omega)$ be its inverse. Then, if $u_1(z)=u(f_\Omega(z))$, $u(w)=u_1(g_\Omega(w))$, then (L1) is equivalent to

$$\begin{aligned}\Delta u_1&=0, &&z\in U,\\ u_1&=\phi_0, &&z\in\partial U,\end{aligned} \tag{L2}$$

where $\phi_0(z)=\phi(f_\Omega(z))$. By equivalent we mean that we may solve (L1) by solving (L2), although ϕ_0 is not unique since it depends on the choice of f_Ω. Now $u_1(z)=\ln|f_\phi(z)|^2$ solves (L2), where f_ϕ is as above except that ϕ is replaced by ϕ_0 in the formula for h. Thus, the solution of (L1) is given by $u(w)=\ln|f_\phi(g_\Omega(w))|=\mathrm{Re}\,h(g_\Omega(w))$, and we may take $v(w)=(f_\phi(g_\Omega(w)))^{-2}$, $w\in\Omega$. As before, it is not the case that $v\in A_N(\Omega)$ for continuous ϕ and smooth $\partial\Omega$. However, (L1) conformally smooth implies $v\in A_N(\Omega)$.

For (P) we only consider v satisfying:

(C2) $v(w)\in A(\Omega)$ has at most simple zeros in Ω and if for $w_0\in\Omega$, $v(w_0)=0$, then $v''(w_0)=0$.

If there exists v satisfying (C1) and (C2), then we say that (P) is solvable via the general integral.

Clearly it is the allowance of simple zeros of v in Ω that leads to the possibility of large norm solutions. If for a given ϕ and Ω we assume that v depends continuously on λ in a neighborhood of $\lambda=0$ and that the zeros of v are independent of λ, then u increases without bound exactly at the zeros of v as $\lambda\to 0_+$.

In order to develop asymptotic solutions of (P) it will be convenient to require, in addition to (C1) and (C2), that v satisfy:

(C3) $v(w;\lambda)\neq 0$ for $w\in\partial\Omega$, where $\lambda\in[0,\lambda^*]$, $\lambda^*>0$; $v(w;\lambda)$ is continuous on $\overline{\Omega}\times[0,\lambda^*]$; the zeros of $v(w;\lambda)$ in Ω are independent of λ; and $v(w;\lambda)$ can be expanded in a power series about $\lambda=0$ that is, $v(w;\lambda)=\sum_{i=1}^{\infty}\lambda^i v_i(w)$.

We say that (P) is power series solvable via the general integral if there exists v satisfying (C1), (C2) and (C3).

3. Equivalent formulations of (P). In this section we discuss equivalent formulations of (P) (in the sense described in the previous section) via conformal transplantation. If $f_\Omega\in H(U)$ is a homeomorphism of $\overline{U}$ onto $\overline{\Omega}$, then (P) is equivalent to

$$\begin{aligned}-\Delta u_1&=\lambda|f_\Omega'|^2e^{u_1}, &&|z|<1,\\ u_1&=\phi_0, &&|z|=1,\end{aligned} \tag{P0}$$

with general integral

$$e^{-u_1(z)/2}=|v_1(z)|^2+\frac{\lambda}{8}\left|v_1(z)\int^z \frac{f_\Omega'(\hat z)\,dz}{(v_1(\hat z))^2}\right|^2,$$

where

$$u_1(z)=u(f_\Omega(z)),\quad \phi_0(z)=\phi(f_\Omega(z)),\quad v_1(z)=v(f_\Omega(z)).$$

We note that solutions of (P0) are essentially independent of $f_\Omega(0)$ and the argument of $f_\Omega'(0)$ (i.e., rotations and translations of Ω).

Letting $f_\phi=e^{(1/2)h(z)}$, where $h(z)$ is given by (2.3) with ϕ replaced by ϕ_0 we rewrite (P0) along with two additional formulations:

(P1)
$$\begin{aligned} -\Delta u_1&=\lambda|f_\Omega'|^2e^{u_1}, && |z|<1,\\ u_1&=\ln|f_\phi|^2, && |z|=1,\end{aligned}$$

(P2)
$$\begin{aligned} -\Delta u_2&=\lambda|f_\phi f_\Omega'|^2e^{u_2}, && |z|<1,\\ u_2&=0, && |z|=1,\end{aligned}$$

(P3)
$$\begin{aligned} -\Delta u_3&=\lambda e^{u_3}, && |z|<1,\\ u_3&=\ln|f_\phi f_\Omega'|^2, && |z|=1,\end{aligned}$$

where

$$u_2=u_1-\ln|f_\phi|^2,\quad u_3=u_1+\ln|f_\Omega'|^2,\quad u_3=u_2+\ln|f_\phi f_\Omega'|^2.$$

with corresponding general integrals

$$e^{-u_1/2}=|v_1|^2+\frac{\lambda}{8}\left|v_1\int\frac{f_\Omega'\,dz}{v_1^2}\right|^2,$$

$$e^{-u_2/2}=|v_2|^2+\frac{\lambda}{8}\left|v_2\int\frac{f_\Omega' f_\phi\,dz}{v_2^2}\right|^2,$$

$$e^{-u_3/2}=|v_3|^2+\frac{\lambda}{8}\left|v_3\int\frac{dz}{v_3^2}\right|^2,$$

where

$$\begin{aligned} v_1(z)&=v(f_\Omega(z)),\\ v_2(z)&=v(f_\Omega(z))(f_\phi(z))^{1/2},\\ v_3(z)&=v(f_\Omega(z))(f_\Omega'(z))^{-1/2}.\end{aligned}$$

Since ϕ continuous and $\partial\Omega$ smooth implies $\ln|f_\phi|$ is continuous on $\overline{U}$, we have the equivalence of (P1) and (P2) under this mild hypothesis; although it is not true that $v_2\in A(U)$ exactly when v_1 is. However, asymptotic solutions of (P2) which are in $C(\overline{U})$ will yield asymptotic solutions of (P) in $C(\overline{\Omega})$. Hence we proceed only with (P2) and modify (C1), (C2), (C3) and the definition of power series solvable to apply to v_2 rather than v. Note that if $f_\Omega', f_\phi\in A_N$ (e.g., if (P) is conformally smooth), then all equivalence difficulties are obviated.

4. Nonuniqueness of the associated map. In this section we consider the nonuniqueness of $f'=f'_\Omega f_\phi$. For $|\tau|<1$ the functions

$$\psi_\tau(z)=\frac{z-\tau}{1-\bar{\tau}z}, \qquad \psi_{-\tau}(z)=\frac{z+\tau}{1+\bar{\tau}z}$$

are inverse conformal maps of $\overline{U}$ onto $\overline{U}$. If $f_{\Omega,0}=f_\Omega\in A(U)$ is a given conformal map of U onto Ω, then we define a family of such conformal maps by

$$f_{\Omega,\tau}=f_{\Omega,0}\circ\psi_{-\tau}.$$

Furthermore if we define $z_0=z$ and $z_\tau=\psi_\tau(z_0)$, then

$$f_{\Omega,\tau}(z_\tau)=f_{\Omega,0}(\psi_{-\tau}(z_\tau))=f_{\Omega,0}(z_0).$$

Following this notation we define

$$\phi_\tau(z_\tau)=\phi(f_{\Omega,\tau}(z_\tau))=\phi(f_{\Omega,0}(z_0))=\phi_0(z_0)$$

and

$$h_\tau(z_\tau)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{it_\tau}+z_\tau}{e^{it_\tau}-z_\tau}\phi_\tau(e^{it_\tau})\,dt_\tau+ic_\tau.$$

It can be shown that $h_\tau(z_\tau)=h_0(z_0)$ exactly when we take

$$c_\tau=c_0+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{2\,\mathrm{Im}(\tau e^{it_0})}{\left(1+|\tau|^2-2\,\mathrm{Re}(\tau e^{-it_0})\right)}\phi_0(e^{it_0})\,dt_0,$$

and we always choose c_τ in this manner. Hence we may define

$$f_{\phi,\tau}(z_\tau)=e^{(1/2)h_\tau(z_\tau)}=e^{(1/2)h_0(z_0)}=f_{\phi,0}(z_0).$$

Now let f_0 satisfy $f_0'=f'_{\Omega,0}f_{\phi,0}$ and define $f_\tau=f_0\circ\psi_{-\tau}$ so that $f_\tau'(z_\tau)=f'_{\Omega,\tau}(z_\tau)f_{\phi,\tau}(z_\tau)$.Thus from a characterization f_0 ($f_0'=f'_{\Omega,0}f_{\phi,0}$) of a region Ω with boundary data ϕ, we construct a family of characterizations via

$$f_{\Omega,\tau}=f_{\Omega,0}\circ\psi_{-\tau},\quad f_{\phi,\tau}=f_{\phi,0}\circ\psi_{-\tau},\quad f_\tau=f_0\circ\psi_{-\tau}$$

with $f_\tau'=f'_{\Omega,\tau}f_{\phi,\tau}$. Any of these characterizations is acceptable when formulating (P2).

Note that $f_0', f_\tau'\in A_N$ if $f'_{0,\Omega}, f_{0,\phi}$ are. Also since solutions of (P2) are independent of $f_\tau(0)$ and the argument of $f_\tau'(0)$ we may always reformulate to obtain $f_\tau(0)=0$ and $f_\tau'(0)>0$.

5. A special case. In this section we consider the special case where Ω is the unit circle with zero boundary data. Hence we take $f'\equiv 1$ ($f_\Omega(z)=z, f_\phi(z)=1$) so that (P2) becomes

(P4)
$$\begin{aligned} -\Delta u&=\lambda e^u, && |z|<1,\\ u&=0, && |z|=1, \end{aligned}$$

with general integral

$$e^{-u/2}=|v|^2+\frac{\lambda}{8}\left|v\int\frac{dz}{v^2}\right|^2,$$

where v has only simple zeros and if $v(z_0)=0$, then $v''(z_0)=0$. We make two choices for v and the integration constant:

$$v_0(z)=\frac{1}{a}, \qquad c_0=0,$$

$$v_1(z)=\frac{z}{a}, \qquad c_1=0.$$

Thus

$$e^{-u_0/2}=\frac{1+(\lambda/8)|\int^z a^2\,d\hat{z}|^2}{|a|^2}=\frac{1+(\lambda/8)|a|^4|z|^2}{|a|^2},$$

$$e^{-w_1/2}=\frac{|z|^2+(\lambda/8)|z\int^z(a^2/\hat{z})\,d\hat{z}|^2}{|a|^2}=\frac{|z|^2+(\lambda/8)|a|^4}{|a|^2}$$

are solutions to (P4) provided $|a|^2$ satisfies the boundary condition which requires that $|a|^2=(4/\lambda)(1+\sqrt{1-\lambda/2})$. It can be shown that changing the sign of the radical simply interchanges the solutions and we lose nothing by taking

$$|a|^2=\frac{4}{\lambda}\left(1-\sqrt{1-\frac{\lambda}{2}}\right)=1+\frac{\lambda}{8}+\frac{\lambda^2}{16}+\cdots.$$

Thus it is clear that a may be taken to be holomorphic in λ for both solutions; thus for more general ϕ and $\partial\Omega$ we expect (P2) to be power series solvable with both the minimal and large norm solutions obtainable in this manner. However in this paper we consider only asymptotic solutions.

6. Asymptotic solutions for (P2). For convenience we rewrite (P2) as

(P5)
$$\begin{aligned} -\Delta u&=\lambda|f'|^2e^u, && z\in U,\\ u&=0, && z\in\partial U,\end{aligned}$$

with general integral

$$e^{-u/2}=|v|^2+\frac{\lambda}{8}\left|v\int\frac{f'}{v^2}\,dz\right|^2, \tag{6.1}$$

where v has only simple zeros in U and if $v(z_0)=0$, $z_0\in U$, then

$$\frac{v''(z_0)}{v'(z_0)}=\frac{f''(z_0)}{f'(z_0)}. \tag{6.2}$$

For (P5), (C1) implies the boundary condition

$$1=|v|^2+\frac{\lambda}{8}\left|v\int\frac{f'}{v^2}\,dz\right|^2 \tag{6.3}$$

and (C2) requires $v\in A(U)$ and (6.2).

To implement (C3) we define

$$v(z)=\frac{v_0(z)}{G(z;\lambda)},$$

where $v_0 \in A$, $G \in A_N$ and $G(z;\lambda)$ is to be expanded in a power series in λ. Hence the general integral (6.1) becomes

$$e^{-u/2} = \frac{|v_0(z)|^2 + (\lambda/8)\left|v_0(z)\int^z \left[(G(\hat{z};\lambda))^2/(v_0(\hat{z}))^2\right] f'(\hat{z})\,d\hat{z}\right|^2}{|G(z;\lambda)|^2}. \tag{6.4}$$

Since we are to solve the boundary condition asymptotically we take $G(z;\lambda)$ to be a finite series:

$$G(z;\lambda) = 1 + \lambda G_1(z) + \cdots + \lambda^{n-1} G_{n-1}(z).$$

Thus we require:

$$|G(z;\lambda)|^2 = |v_0(z)|^2 + \left| v_0(z) \int^z \frac{[G(\hat{z};\lambda)]^2 f'(\hat{z})}{[v_0(\hat{z})]^2}\,dz \right|^2, \qquad z \in \partial U. \tag{C1'}$$

(C2′) $v_0 \in A$ has at most simple zeros in U. If for $z_0 \in U$ we have $v_0(z_0) = 0$, then

$$\frac{v_0''(z_0)}{v_0'(z_0)} = \frac{f''(z_0)}{f'(z_0)}$$

and

$$G_i'(z_0) = 0, \qquad i = 1, \cdots, n.$$

(C3′) $G_i \in A$, $i = 1, \cdots, n$ and $v_0(z) \neq 0$ for $z \in \partial U$.

First order conditions (which involve only v_0) are obvious and are given in the next section. Detailed computations for G_i are given in [14] and summarized in §8.

7. First order solutions of (P5). Implementing (C1′)–(C3′) to first order yields:

(FC1) $1 = |v_0(z)|^2$ for $z \in \partial U$.

(FC2a) $v_0 \in A$ has simple zeros in U.

(FC2b) If $v_0(z_0) = 0$, then $v_0''(z_0)/v_0'(z_0) = f''(z_0)/f'(z_0)$.

(FC3) $v_0(z) \neq 0$ for $z \in \partial U$.

Conditions (FC1), (FC2a) and (FC3) imply that v_0 is a finite Blaschke product:

$$v_{0p}(z) = k_0 \prod_{i=1}^{p} \frac{z - \delta_i}{1 - \bar{\delta}_i z}, \qquad p = 1, 1, 2, \cdots,$$

where $|k_0| = 1$ and $\delta_i \neq \delta_j$ for $i \neq j$. Since u is independent of the choice of k_0, we take $k_0 = 1$ (thus $v_{00} \equiv 1$). The condition (FC2b) is left as a constraint on f for $p \geq 1$. Specifically, at the zeros of v_0 $(\delta_i, i = 1, \cdots, p)$ we require

$$\frac{f''(\delta_i)}{f'(\delta_i)} = C_i(\delta_i, \cdots, \delta_p), \qquad i = 1, \cdots, p, \tag{7.1}$$

where

$$C_i(\delta_1, \cdots, \delta_p) = \frac{2\bar{\delta}_i}{1 - |\delta_i|^2} + 2 \sum_{\substack{j=1 \\ j \neq 1}}^{p} \frac{1 - |\delta_j|^2}{(1 - \bar{\delta}_j \delta_i)(\delta_i - \delta_j)}.$$

Examples of associated maps f (with zeros δ_i, $i=1,\cdots,p$) satisfying (7.1) for $p\geq 2$ (e.g., multiple maxima for $\phi\equiv 0$) are not in evidence. Hence for the remainder of this paper we consider only $p=0,1$ (the minimal and large norm solutions). For $p=1$ let

$$v_{01}(z)=\frac{z-\delta}{1-\bar{\delta}z}$$

so that (7.1) becomes

$$\frac{f''(\delta)}{f'(\delta)}=\frac{2\bar{\delta}}{1-|\delta|^2} \tag{7.2}$$

(the first order constraint of Weston, see (1.1)).

For $f'\in A_N$, (7.2) has at least one solution. To investigate how a solution δ varies with the choice of associated map f (see §4) we consider a solution of (7.2) to be an ordered pair (f,δ). It is a straightforward computation to show that if (f_0,δ_0) satisfies (7.2) then (f_τ,δ_τ), where $f_\tau=f_0\circ\psi_{-\tau}$ and $\delta_\tau=\psi_{-\tau}(\delta_0)$ does also. Furthermore, since $f_\tau(\delta_\tau)=f_0(\delta_0)$ for all $\tau\in U(f_{\delta_0},0)$ satisfies (7.2); that is $f_{\delta_0}''(0)=0$.

To simplify further computation of the asymptotic solution as well as the modified Newton's method, if (f,δ) satisfies (7.2) we normalize f by

$$f_N(z)=e^{i\beta}\left[f(\psi_{-\delta}(z))-f(\psi_{-\delta}(0))\right], \tag{7.3}$$

where β is the argument of $d(f(\psi_{-\delta}(z)))/dz$ evaluated at zero. Hence $f_N(0)=0$ and $f_N'(0)>0$ as well as $f_N''(0)=0$.

We consider some examples:

(EX1)
$$f(z)=\frac{k(z-z_1)}{1-z_0z},\quad |z_0|<1,\quad z_1z_0\neq 1,$$
$$\delta=\bar{z}_0,\quad f_N(z)=k_1z,\quad k_1=|k|\frac{|1-z_1z_0|}{1-|z_0|^2}.$$

(EX2)
$$f(z)=k(z+\alpha z^2),\qquad |\alpha|<\frac{1}{2},$$
$$\delta=\bar{\alpha}\left(\frac{\sqrt{1+12|\alpha|^2}-1}{6|\alpha|^2}\right),\qquad f_N(z)=k_1\frac{z(1+2k_2z)}{(1+k_2z)^2},$$
$$k_1=\frac{|k|\left(1-|k_2|^2\right)^2}{1-3|k_2|^2},\qquad k_2=\alpha\left(\frac{\sqrt{1+12|\alpha|^2}-1}{6|\alpha|^2}\right).$$

(EX3)
$$f(z)=k\left(\frac{e^{\alpha z}-1}{\alpha}\right),\qquad 0\leq|\alpha|<\pi,$$
$$\delta=\frac{\bar{\alpha}\left(\sqrt{1+|\alpha|^2}-1\right)}{|\alpha|^2},\qquad f_N(z)=\frac{k_1}{k_2}\left(\exp\left\{\frac{2k_2z}{1-k_2z}\right\}-1\right),$$
$$k_1=\frac{|k|\left(1-|k_2|^2\right)}{|\alpha|^2}\exp\left\{\frac{2|k_2|^2}{1-|k_2|^2}\right\},\qquad k_2=\frac{\alpha\left(\sqrt{1+|\alpha|^2}-1\right)}{|\alpha|^2}.$$

(EX4) $$f(z)=k(z+\alpha z^3), \qquad |\alpha|<\frac{1}{3},$$
$$\delta=0, \qquad f_N(z)=k(z+\alpha z^3).$$

(EX5) $$f(z)=\frac{kz}{1+\alpha z^2}, \qquad |\alpha|<1,$$
$$\delta=0, \qquad f_N(z)=\frac{|k|z}{1+\alpha z^2}.$$

We note that for all of the above f is in fact univalent on $\overline{U}$.

(EX6) If $f''(0)=0$ (e.g., if f is odd), then we may take $\delta=0$.

(EX7) Assume zero boundary data and take $f=f_\Omega$. If Ω is symmetric with respect to two perpendicular axes which we may choose to be the real and imaginary axes in the w-plane, we can construct (by the Schwarz reflection principle) f: $\overline{U}\to\overline{\Omega}$ which is odd. Hence take $\delta=0$.

(EX8) Assume zero boundary and take $f=f_\Omega$. Let Ω be symmetric with respect to a line which we may choose to be the real axis in the w-plane. We construct f (by the Schwarz reflection principle) such that $f(x)$ is real for x real. We look for a real solution δ to (7.2) at the intersection of the curves $f''(x)/f'(x)$ and $2x/(1-x^2)$ on the interval $(-1,1)$. If $f''(x)/f'(x)$ is bounded, the curves must intersect at least once.

For the remainder of the paper we drop the subscript on f and always assume that f is normalized; that is

(7.4) $$f''(0)=0, \quad f(0)=0 \quad \text{and} \quad f'(0)>0.$$

Using (7.4) and (6.4) we obtain the first order asymptotic solutions

(7.5) $$e^{-u_{10}/2}=1+\frac{\lambda}{8}|f(z)+c_{00}|^2 \quad \text{(minimal)},$$

(7.6) $$e^{-u_{11}/2}=|z|^2+\frac{\lambda}{8}|-f'(0)+zI_0(z)+c_{01}z|^2 \quad \text{(large norm)},$$

where

(7.7) $$I_0(z)=\int_0^z (f'(\hat{z})-f'(0))\,\frac{d\hat{z}}{\hat{z}^2}$$

and c_{00} and c_{01} are constants of integration. If $f'\in A_N$ then it can easily be shown that $I_0(z)$ is continuous on $\overline{U}$, in fact, $I_0(z)\in A_{1/2}$.

THEOREM 2. *If $f'\in A_N$ satisfies* (7.4), *then* (7.5) *and* (7.6) *define first order asymptotic solutions of* (P5) *which are in $C(\overline{U})\cap C^2(U)$, satisfy the partial differential equation*

$$-\Delta u=\lambda|f'(z)|^2e^u$$

exactly in U and satisfy the boundary condition

$$u=0 \quad \text{on } |z|=1$$

to order 1 *as* $\lambda\to 0$; *that is*

$$\max_{|z|=1}|u_{1p}|=O(\lambda), \qquad p=0,1.$$

Furthermore

$$\max_{|x|\le 1}|u_{10}| = O(\lambda) \quad (\textit{minimal}),$$

$$\max_{|x|\le 1}|u_{11}| = O(\ln(1/\lambda)) \quad (\textit{large norm single maximum}).$$

8. Second and higher order results for (P5). In this section we consider second and higher order asymptotic solutions for the minimal and large norm single maximum cases ($p=0,1$). We illustrate the procedure using the second order case. Detailed computations are given in [14].

The second order requirements on G_{1p} are given by:

(SC1) $$2\operatorname{Re} G_{10} = \frac{1}{8}\left\{|c_{00}|^2 + 2\operatorname{Re}(\bar{c}_{00}f(z)) + |f(z)|^2\right\} \quad \text{on } |z|=1 \quad \text{(minimal)},$$

$$2\operatorname{Re} G_{11} = \frac{1}{8}\left\{|c_{01}|^2 + 2\operatorname{Re}(-f'(0)c_{01}z + \bar{c}_{01}I_0(z)) + |f'(0) + zI_0(z)|^2\right\}$$

$$\text{on } |z|=1 \quad \text{(large norm)}.$$

(SC2) $$G'_{11}(0) = 0 \quad \text{(large norm)}.$$

(SC3) $$G_{10} \in A(U) \quad \text{(minimal)},$$

$$G_{11} \in A(U) \quad \text{(large norm)}.$$

Since $I_0 \in A_{1/2}(U)$

$$\mathcal{S}_{00}(z) = |f(z)|^2, \qquad \mathcal{S}_{01}(z) = |-f(0) + zI_0(z)|^2$$

are both continuous on ∂U, and in fact, for $p=0,1$ we have $\mathcal{S}_{0p}(e^{i\theta}) \in \Lambda_{1/2}$, where $\Lambda_\alpha = \{q\colon \mathbb{R}\to\mathbb{R}$, where q is periodic of period 2π and satisfies a Lipschitz condition of order $\alpha\}$. Hence using Schwarz's formula we obtain:

(8.1)

$$G_{10}(z) = \frac{1}{8}\left\{\frac{|c_{00}|^2}{2} + \bar{c}_{00}f(z) + \frac{1}{2\pi i}\int_{|\hat{z}|=1}\mathcal{S}_{00}(\hat{z})\left(\frac{1}{\hat{z}-z} - \frac{1}{2\hat{z}}\right)d\hat{z}\right\} \quad \text{(minimal)},$$

(8.2)

$$G_{11}(z) = \frac{1}{8}\left\{\frac{|c_{01}|^2}{2} - \bar{f}'(0)c_{01}z + \bar{c}_{01}I_0(z) + \frac{1}{2\pi i}\int_{|\hat{z}|=1}\mathcal{S}_{01}(\hat{z})\frac{1}{\hat{z}-z} - \frac{1}{2\hat{z}}\,d\hat{z}\right\} \quad \text{(large norm)}$$

with $G_{1p} \in A_{1/2}$ for $p=0,1$. To satisfy (SC2) we require that

$$c_{01} = \frac{f'(0)Z_{01} + (f''(0)/2)Z_{01}}{\det},$$

where

$$Z_{01} = \frac{1}{2\pi i}\int_{|\hat{z}|=1}\mathcal{S}_{01}(z)\,\frac{d\hat{z}}{\hat{z}^2},$$

$$\det = |f'(0)|^2 - \frac{1}{4}|f'''(0)|^2.$$

Hence we require

$$|f'''(0)| \neq 2|f'(0)| \tag{8.3}$$

(the simplified second order constraint of Weston, see (1.3)). We can thus write the second order solutions as

$$e^{-u_{20}/2} = \frac{1+(\lambda/8)|c_{00}+f(z)+\mathcal{Q}_0|^2}{|1+\lambda G_{10}(z)|^2} \tag{8.4}$$

with $\mathcal{Q}_0 = \lambda(c_{10}+I_{10}(z))+\lambda^2 \int^z G_{10}^2(\hat{z}) f'(\hat{z})\, d\hat{z}$ and

$$e^{-u_{21}/2} = \frac{|z|^2+(\lambda/8)|-f'(0)+c_{01}z+zI_0(z)+\mathcal{Q}_1|^2}{|1+\lambda G_1(z)|^2} \tag{8.5}$$

with $\mathcal{Q}_1 = \lambda(-2G_0(0)f'(0)+c_{11}z+zI_{11}(z))+\lambda^2 z^2 \int^z (G_{11}^2(\hat{z}) f'(\hat{z})\, d\hat{z}/\hat{z}^2)$, where

$$I_{10}(z) = \int_0^z 2G_{10}(\hat{z}) f'(\hat{z})\, d\hat{z},$$

$$I_{11}(z) = \int_0^z (2G_{11}(\hat{z}) f(\hat{z}) - 2G_{11}(0) f'(0)) \frac{d\hat{z}}{\hat{z}^2}$$

and c_{10} and c_{11} are arbitrary integration constants. Recall that c_{01} is determined above but c_{00} and all higher order (λ^2) integration constants are arbitrary.

The higher order requirements for G_{ip}, $2 \leq i \leq n-1$, $p=0,1$, are similar to those for G_{ip}. The constants c_{i1}, $1 \leq i \leq n-2$ are determined similar to c_{01} and no further requirements, other than (8.3), are needed to satisfy $G_i'(0)=0$ for $2 \leq i \leq n$. Also G_{ip} can be shown to be in $A_{1/2}$. The explicit formulas for G_i and c_{i-1} for $2 \leq i \leq n-1$ are given in [14]. Hence we obtain the nth order asymptotic solutions:

$$e^{-u_{np}/2} = \frac{|v_{0p}|^2 + \frac{\lambda}{8}\left|\sum_{i=0}^{n-1} \lambda^i M_{ip} + \sum_{i=n}^{2(n-1)} \lambda^i M_{ip}^n\right|^2}{\left|1+\sum_{i=1}^{n-1} \lambda^i G_{ip}\right|^2}, \qquad p=0,1, \tag{8.6}$$

where for $i=2,\cdots,n-1$ and $p=0,1$

$$v_{00}(z)=1,$$

$$v_{01}(z)=z,$$

$$G_{0p}(z)=1,$$

$$M_{i0}(z)=c_{i0}+I_{i0}(z),$$

$$M_{i1}(z)=-f'(0)L_{i1}(0)+c_{i1}z+zI_{i1}(z),$$

$$M_{i0}^n(z)=\int^z L_{i0}^n(\hat{z}) f'(\hat{z})\, d\hat{z},$$

$$M_{i1}^n(z)=z\int^z L_{i1}^n(\hat{z}) f'(\hat{z}) \frac{d\hat{z}}{\hat{z}^2},$$

$$I_{i0}(z)=\int_0^z L_{i0}(\hat{z}) f'(\hat{z})\, d\hat{z},$$

$$I_{i1}(z)=\int_0^z (L_{i1}(\hat{z}) f'(\hat{z}) - L_{i1}(0) f'(0)) \frac{d\hat{z}}{\hat{z}^2},$$

$$L_{ip}(z)=\sum_{j=0}^{i} G_j(z)G_{i-j}(z),$$

$$L_{ip}^{n}(z)=\sum_{j=i-(n-1)}^{n-1} G_j(z)G_{i-j}(z),$$

$$G_{i0}(z)=\frac{1}{8}\left\{c_{01}\bar{c}_{i-1,0}+\frac{1}{2\pi i}\int_{|\hat{z}|=1}\mathbb{S}_{i-1,0}(\hat{z})\left(\frac{1}{\hat{z}-z}-\frac{1}{2\hat{z}}\right)d\hat{z}\right\},$$

$$G_{i1}(z)=\frac{1}{8}\left\{c_{01}\bar{c}_{i-1,1}-\overline{f'(0)}c_{i-1,1}z+I_0(z)\bar{c}_{i-1,1}\right.$$

$$\left.+\frac{1}{2\pi i}\int_{|\hat{z}|-1}\mathbb{S}_{i-1,1}(\hat{z})\left(\frac{1}{\hat{z}-z}-\frac{1}{2\hat{z}}\right)d\hat{z}\right\}$$

and for $i=2,\cdots,n-2$ and $p=0,1$

$$\mathbb{S}_{ip}(z)=\mathbb{f}_{ip}(z)+\mathbb{R}_{ip}(z),$$

$$\mathbb{f}_{i0}(z)=2\,\mathrm{Re}[M_{00}(z)I_{i0}(z)]=2\,\mathrm{Re}[(c_{00}+f(z))\overline{I_{i0}(z)}],$$

$$\mathbb{f}_{i1}(z)=2\,\mathrm{Re}[M_{01}(z)(-f'(0)L_{i1}(0)+I_{i1}(z))],$$

$$\mathbb{R}_{ip}(z)=-8\sum_{j=1}^{i} G_{jp}(z)\bar{G}_{i-j+1,p}(z)+\sum_{j=1}^{i-1} M_{jp}(z)M_{i-j,p}(z).$$

We summarize our results in the following:

THEOREM 3. *If $f'\in A(U)$ and satisfies* (7.4) *and* (8.3), *then* (8.6) *defines nth order asymptotic solutions of* (P5) *which are in $C(\bar{U})\cap C^2(U)$, satisfy*

$$-\Delta u=|f'(z)|^2e^u$$

exactly in U and satisfy

$$|u|=0 \quad \textit{on } |z|=1$$

to order n as $\lambda\to 0$; that is

$$\max_{|z|=1}|u_{np}|=O(\lambda^n), \qquad p=0,1.$$

Furthermore,

$$\max_{|z|\le 1}|u_{n0}|=O(\lambda), \qquad \max_{|z|\le 1}|u_{n1}|=O(\ln(1/\lambda))$$

as $\lambda\to 0$.

Straighforward computations show that for (EX1)–(EX4) the normalized functions f_N satisfy (8.3) as well as (7.4). However for (EX5)–(EX8) (8.3) remains as an additional constraint.

9. A modified Newton's method for (P5). In this section we show that the large norm asymptotic solution for (P5) given in the previous section, (8.6) with $p=1$, has the ability to generate an exact solution via a modified Newton's iteration scheme provided that the asymptotic solution is taken to order n ($n\ge 3$) and that λ is sufficiently small.

To convert (P5) to the equivalent integral equation ($u\in C(\bar{U})$ with $\|u\|=\max_{|x|\le 1}|u(x)|$, $x=(x_1,x_2)$, $dx=dx_1dx_2$)

(P6) $$u=\mathbb{K}u,$$

where

$$(\mathbb{K}u)(x_0)=\int_{|x|\le 1} g(x_0,x)e^{u(x)}|f'(z)|^2\,dx$$

and

$$g(x_0,x)=\frac{1}{4\pi}\ln\left|\frac{z-z_0}{1-z_0 z}\right|^2$$

is the Green's function for the Dirichlet problem associated with the unit disk. Here x and x_0 are two-dimensional vectors with components given by the real and imaginary parts of the complex numbers z and z_0 respectively. Note that since the singularity of $g(x_0,x)$ is weakly polar that $\mathbb{K}$ maps $C(\bar{U})$ into $C(\bar{U})$.

The modified Newton's method

$$u_{n+1}=\mathbb{S}(u_n), \tag{9.1}$$

where

$$\mathbb{S}(u)=\left(I-\mathbb{K}'_{u_0}\right)^{-1}\left(\mathbb{K}(u)-\mathbb{K}'_{u_0}(u)\right),$$

$$\left(\mathbb{K}'_{u_0}h\right)(x_0)=\int k(x_0,x)h(x)\,dx$$

and

$$k(x_0,x)=\lambda g(x_0,x)e^{u_0(x)}|f'(x)|^2$$

(i.e., $\mathbb{K}'_{u_0}$ is the Fréchet derivative of $\mathbb{K}$ evaluated at u_0) may be developed in the same manner as in [17] to yield:

THEOREM 4. *If* $\|u_0-\mathbb{K}(u_0)\|\le\ln(1+(1/\Gamma))-1/(1-\Gamma)$, *where*

$$\|\left(I-\mathbb{K}'_{u_0}\right)^{-1}\mathbb{K}'_{u_0}\|\le\Gamma,$$

then the modified Newton's method (9.1) *will converge to a unique solution* u^* *of* (P6) *such that*

$$\|u-u^*\|<\ln\left(1+\frac{1}{\Gamma}\right).$$

Under the conditions of Theorem 3 we take the large norm asymptotic solution, (8.6) with $p=1$, as the initial approximation u_0 in (9.1). By following the procedure of [17] we easily have

$$\|u_0-\mathbb{K}u_0\|\le M_1\lambda^n$$

for λ sufficiently small, where n is the order of the asymptotic solution. Furthermore, the methods of [17] (also see [14]) yield

$$\|\mathbb{K}'_{u_0}\|\le M_2\ln\left(\frac{1}{\lambda}\right),\qquad \|\left(I-\mathbb{K}'_{u_0}\right)^{-1}\|\le M_3\lambda^{-1}$$

and hence

$$\Gamma\le M_4\frac{1}{\lambda}\ln\left(\frac{1}{\lambda}\right)$$

for λ sufficiently small. Now for $\Gamma>1$

$$\ln\left(1+\frac{1}{\Gamma}\right)-\frac{1}{1+\Gamma}=\sum_{n=1}^{\infty}(-1)^{n+1}\left(\frac{1}{n}-1\right)\left(\frac{1}{\Gamma}\right)^n$$

$$=\frac{1}{2}\left(\frac{1}{\Gamma}\right)^2-\frac{2}{3}\left(\frac{1}{\Gamma}\right)^3+\cdots.$$

Hence if λ is sufficiently small we may apply Theorem 4 provided

$$\|u_0-K(u_0)\|\leq M_5\lambda^2\left(\ln\left(\frac{1}{\lambda}\right)\right)^{-2}.$$

This is attained if $n\geq 3$ and λ is sufficiently small. Thus we have:

THEOREM 5. *Let the conditions of Theorem* 3 *hold and the large norm asymptotic solution* (8.6) *with* $p=1$ *and* $n\geq 3$ *be used in the modified Newton's method* (9.1). *Then* u_n *will converge to a unique large norm solution* u^* *of* (P6) *such that*

$$\|u_0-u^*\|=O(\lambda).$$

Clearly u^* *is also an exact solution of* (P5). *Furthermore if* $f'=f_\Omega' f_\phi$, *where* f_Ω *and* f_ϕ *are as in* §1 *with* g_Ω *the inverse mapping for* f_Ω, *then* $u^*\circ g_\Omega$ *is an exact solution of* (P2) *and* $u^*\circ g_\Omega+\ln|f_\phi\circ g_\Omega|$ *is an exact solution of* (P).

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1966.

[2] C. BANDLE, *Existence theorems, qualitative results and a priori bounds for a class of nonlinear Dirichlet problems*, Arch. Rat. Mech. Anal., 58 (1975), pp. 219–238.

[3] ———, *Isoperimetric inequalities for a nonlinear eigenvalue problem*, Proc. Amer. Math. Soc., 56 (1976), pp. 243–246.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, *Vol.* II, Interscience, New York, 1962.

[5] M. G. CRANDALL AND P. N. RABINOWITZ, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Rat. Mech. Anal., 58 (1975), 207–218.

[6] I. M. GELFAND, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Trans., 29 (1963), pp. 295–381.

[7] B. GIDAS, WEI-MING NI AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[8] G. M. GOLUZIN, *Geometric Theory of Functions of a Complex Variable*, American Mathematical Society, Providence, RI, 1969.

[9] J. L. KAZDAN AND F. W. WARNER, *Curvature functions for compact* 2*-manifolds*, Ann. Math., 99 (1974), pp. 14–47.

[10] J. B. KELLER AND D. S. COHEN, *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), pp. 1361–1376.

[11] J. LIOUVILLE, *Sur l'équation aux derivées partielles* $(\partial^2\log\lambda)/\partial u\partial v\pm 2\lambda a^2=0$, J. de Math., 18 (1953), pp. 71–72.

[12] D. LONDON, *On the zeros of the solutions of* $w''(z)+p(z)w(z)=0$, Pacific J. Math., 12 (1962), pp. 979–991.

[13] C. B. MORREY, *On the analyticity of the solutions of analytic nonlinear elliptic systems of partial differential equations*, I *and* II, Amer. J. Math., 80 (1958), pp. 198–237.

[14] J. L. MOSELEY, *On asymptotic solutions for a Dirichlet problem with an exponential nonlinearity*, AMR-1, West Virginia Univ., Morgantown, 1981.

[15] Z. NEHARI, *Conformal Mapping*, McGraw-Hill, New York, 1952.

[16] C. POMMERENKE, *Univalent Functions*, Vanderhoeck & Ruprecht, Gottingen, 1975.

[17] V. H. WESTON, *On the asymptotic solution of a partial differential equation with an exponential nonlinearity*, this Journal, 9 (1978), pp. 1030–1053.

[18] A. ZYGMUND, *Trigonometric Series*, Cambridge Univ. Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0008 $01.25/0

BIFURCATING INSTABILITY OF THE FREE SURFACE OF A FERROFLUID*

EVAN EUGENE TWOMBLY† AND J. W. THOMAS‡

Abstract. Consider a slab of ferrofluid bounded below by a fixed boundary and above by a vacuum. If the fluid is subjected to a vertically directed magnetic field of sufficient strength, surface waves appear.

The equations which describe this phenomenon are derived. In the physical space no natural Banach space structure is available due to the free surface. In order to use the available bifurcation theory, a transformation of coordinates is made, mapping the surface flat. In the new coordinate system the equations define an operator between Banach spaces. The minimum eigenvalue of the linearized operator is the critical magnetic field strength where the planar surface loses stability.

Using a generalized inverse of the Fréchet derivative of the operator and the implicit function theorem, the existence of a nontrivial branch of solutions is proved. A local stability criterion is also obtained and applied to three periodic structures: rolls, squares and hexagons.

1. Introduction. Fluids with strong magnetic properties have been produced via colloidal suspension of ferromagnetic particles in a suitable carrier fluid. When a horizontal slab of such a "ferrofluid" is in the presence of a vertically directed static magnetic field of sufficient strength, a horizontal plane surface will change. Analogous to Benard cells, both rectangular and hexagonal periodic surface relief patterns have been observed, though the rectangular pattern is rare [1]. In the electrical analogue, a dielectric in an electric field, both lattice structures can be obtained [2]. For a general article on ferrofluids see Moskowitz [3].

Over the past 15 years this and related phenomena have been under investigation. Formal mathematical techniques have produced results in agreement with experimental data. Cowley and Rosensweig [1] develop a set of equations which describe the phenomena and, by assuming linear stability theory [4], are able to find a critical magnetic field strength where the planar surface loses stability. The experimental results they achieve agree with the predicted values. When second order effects are considered, both Gailitis [5] and Kuznetsov and Spector [2] find that the hexagonal surface pattern must jump to a (possibly small) finite height since no hexagonal pattern near zero height is static when the critical magnetic field is exceeded. The same argument produces a static nonplanar solution close to zero when the magnetic field is below the critical field strength, but no stability analysis has been made.

Simplifying the problem to two dimensions, analogous to "rolls" in the Benard problem, Zaitev and Shliomis [6] found static solutions above and below the critical field strength. Which of the two solutions occurs is given as a function of the magnetic permeability of the ferrofluid. Experimentally, rolls can be produced if an additional magnetic field is present. When Barkov and Bashtovoi [7] added a horizontal field, the rolls appeared parallel to the applied field.

In another paper by Gailitis [8] the relative stabilities of the different surface patterns are determined using energy considerations. They asume existence of convergent expansions of hexagonal, square and roll type periodic branches of solutions. The hexagaonal surface pattern was at a lower energy than the other relief patterns, except in the case of a higher magnetic field strength in a ferrofluid of low relative permeability. In that case the rectangular pattern was favored.

Received by the editors January 11, 1982, and in revised form May 26, 1982.
† NHC Wind Engineering, 22477 72nd Ave. S., Kent, Washington 98032.
‡ Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.

Some analyses have been made with the physical setup slightly modified. Bashtovoi [9] let the ferrofluid have both the upper and lower sides of the fluid free to deform. When the thickness of the fluid is small the interactions of the two interfaces affected the critical magnetic field strength. Zelazo and Melcher [10] again used a single free surface but allowed the magnetic field to be oriented at an arbitrary angle. The critical magnetic field strength increased as the horizontal component of the field increased. The above analyses have not been carried out with the mathematical rigor available for analysis of nonlinear phenomena. For a bibliography of papers concerning ferrofluids see [11].

The problem considered in this paper is similar to that considered by [1], [2], [5], [6], [8]. Ferrofluid is placed below a less dense fluid (or vacuum) both having a large finite depth D. The interface between the fluids is restricted only by the physical properties of the fluids. The upper and lower surfaces (at $\pm D$) are flat and any perturbed magnetic field does not extend through the surface.

Through the application of rigorous mathematical techniques we intend to prove that the planar surface loses stability when the critical magnetic field strength is exceeded, and that a new static surface pattern appears at that point. The linear stability associated with the new surface provides conditions for the local stability of this new surface. Note that this stability is the stability of the new surface relative to any other surfaces with identical symmetry properties (i.e., hexagonal, rectangular or rolls), not the stability of one symmetry pattern to another.

The existence and stability proofs will be proved using some standard bifurcation techniques. However, the presence of the free surface complicates the problem. It is for this reason that there are almost no rigorous treatments of nonlinear free surfaces. We remove the difficulties caused by the free surface by using a variation of the method of domain perturbation in Joseph and Fosdick [12], and model our approach after the application of this technique by Sattinger [13]. This method consists of mapping the surface flat and proceeding with the analysis.

The surface patterns to which the stability criterion will be applied are those mentioned above: rolls, rectangles and hexagons. Some general results on these types of patterns [14] indicate that the second-order approximation should be adequate to determine stability of the hexagonal pattern while rolls and rectangles will require third-order approximations.

An outline of the paper is given in the following paragraphs.

Equations must first be obtained which describe the fluid properties of the interface and the interactions of the magnetic fields with the shape of the interface. Since we are looking for static solutions the equations will contain no time derivatives. The constant magnetic permeability provides a continuous magnetostatic potential which satisfies Laplace's equation. Additionally the gradient of the potential when multiplied by the local permeability, $\nabla(\mu\phi)$, must be continuous normally across the interface. An additional condition at the fluid interface is derived which balances the fluid's internal stresses so that a static situation is maintained. A Neumann condition is added at the upper and lower boundaries so the perturbation field does not extend beyond the boundaries.

Under the above conditions, when the surface is planar, a potential can be found. We look for perturbations from this "trivial" solution. This "trivial" solution is subtracted and boundary conditions are set. The expected periodicity of the nonplanar interface allows us to reduce the horizontal domain to a finite region and add Neumann conditions on the sides of the reduced domain. At this point the potential is only

determined to within a constant. A normalizing condition is added and the equations describing the phenomena are complete.

The next step in the process is to apply the domain perturbation technique to the system of equations described above. A transformation is constructed which leaves all the external boundaries fixed but maps the interface to a plane, parallel to the upper and lower surfaces. The normal vector (a contravariant tensor), the gradient (a covariant tensor) and the Laplacian (the divergence of the gradient) are transformed using tensor techniques. When the equations are rewritten in the new coordinate system using the new forms of the normal vector, gradient, and Laplacian, a continuous operator, F, between Banach spaces can be constructed so that solving $F=\mathbf{0}$ is equivalent to solving the desired equation. The Banach space which is the domain of the operator contains many of the linear boundary conditions, and the remaining information is included in the operator, F, itself.

We are finally in a situation where we can apply the standard techniques of bifurcation theory. The first step is to obtain the Fréchet derivative, $F_\phi(\mathbf{0})$, at the trivial solution and find its lowest positive eigenvalue. The linearization of the equations before the transformation and $F_\phi(\mathbf{0})\mathbf{u}=\mathbf{0}$ are the same; hence, comparisons can easily be made with other linearizations. The minimum eigenvalue that is found has been called the critical magnetic field strength, and as D approaches ∞ the critical magnetic field strength that we find approaches the value found by other authors in the case of infinite depth. The associated eigenfunction is also found. The dimension of the null space of $F_\phi(\mathbf{0})$ is one, so the eigenvalue is simple. A linear functional, bounded on the Banach spaces mentioned here, is found which maps the image of $F_\phi(\mathbf{0})$ to zero and will be used later to define a projection to the null space of $F_\phi(\mathbf{0})$.

To obtain stability and existence results similar to those in Sattinger [4], we need a projection and a generalized inverse of $F_\phi(\mathbf{0})$. To this end, a series of lemmas are proved. The generalized inverse depends on the Fredholm alternative and a compact embedding theorem for spaces defined over finite domains. If D were allowed to be infinite the embedding would, in general, not be compact, and other techniques to obtain the generalized inverse would be needed.

The existence of a bifurcating branch of solutions to $F=\mathbf{0}$ extending from the trivial branch is proven as follows. An operator between Banach spaces is defined. The operator takes three components (a parameter, ε, the magnetic field strength as a function of ε, and a vector of the potential and interfacial perturbations as a function of ε) to two components (the image of F projected into the null space of F_ϕ and the generalized inverse of F_ϕ applied to the image of F). When the implicit function theorem is applied to this operator, two results are realized: the existence of the bifurcating branch of solution to $F=\mathbf{0}$ and the local convergence of a power series in ε of the branch.

To obtain stability conditions of the branch, a solution, $\mathbf{v}$, along the branch sufficiently close to the bifurcation point is chosen. If the power series is substituted in F and the coefficients of the powers of ε are set to zero, expressions for the coefficients of the power series are obtained. To relate these coefficients to the conditions of stability in the linear theory we consider $F_\phi(\mathbf{v})\mathbf{u}=\sigma\mathbf{u}$. Perturbations tend to grow when σ is positive and decay for negative σ. By applying the implicit function theorem (§5) to the stability equation, we obtain convergent power series expansions for $\mathbf{u}$ and σ which allows us to determine the sign of σ and hence the stability of the solution $\mathbf{v}$.

The expansion results show that nontrivial branches initially extending towards smaller magnetic field strengths (subcritical) are locally unstable. Branches extending to

greater field strengths (supercritical) are locally stable. When these results are applied to the hexagonal pattern, unstable (subcritical) branches are found. When the approximate stability criterion described earlier is applied to analyzing rolls, the branch is found to be stable (supercritical). For squares, the approximate analysis shows the stability to be a function of permeability. These results are compared with other papers and a summary is given.

2. The mathematical formulation. Our mathematical description of the phenomena is based on Maxwell's equations in electromagnetic theory and on the fact that total stresses and total forces equal zero for a steady state condition.

In this article the fluid has many idealized properties. Since the effect studied is a surface deformation and the ferrofluids are liquid, the fluid is assumed to be incompressible and of large but finite depth, D. After any short-term effects have died out [15] (say, two seconds) and prior to any long-term separation of carrier fluid and suspended particles [16] (say, two hours), it is reasonable to assume that the ferrofluid is magnetically linear, isotropic and free of internal currents when a magnetic field of limited strength is applied [17]. Furthermore, the magnetic permeability is taken to be constant throughout the field.

Under the above assumptions, when all time derivatives are set to zero, the following forms of Maxwell's equations hold:

$$\oint_C \mathbf{H}\cdot\mathbf{t}\,dl=0 \tag{2.1}$$

around a closed curve C where $\mathbf{H}$ is the magnetic field and $\mathbf{t}$ is a unit vector tangent to C,

$$\oint_Q \mathbf{B}\cdot\mathbf{n}\,d\sigma=0 \tag{2.2}$$

over a closed surface Q where $\mathbf{B}$ is the induction field and $\mathbf{n}$ is a unit vector normal to Q, and

$$\tilde{\mu}\mu_0\mathbf{H}=\mathbf{B}, \tag{2.3}$$

where $\tilde{\mu}$ is the relative permeability and μ_0 is the permeability of space.

We choose a coordinate space (y_1,y_2,y_3) so that gravity acts in the negative y_3 direction. Let the interface $y_3=z(y_1,y_2)$ separate the ferrofluid below with constant permeability μ from any other less dense fluid above with constant permeability normalized to one (Fig. 1). Except at the interface and at the upper and lower boundaries, (2.1) and (2.2) are equivalent to the following:

$$\check{\nabla}\times\mathbf{H}=0 \tag{2.4a}$$

and

$$\check{\nabla}\cdot\mathbf{B}=0 \tag{2.4b}$$

where $\check{\nabla}=(\partial/\partial y_1,\partial/\partial y_2,\partial/\partial y_3)^T$, and superscript T indicates vector transpose. Poisson's theorem and (2.4a) guarantee the existence of a scalar potential, ψ, with a negative gradient $\mathbf{H}$. The constant permeability and (2.3) and (2.4b) force the potential to satisfy Laplace's equation:

$$\begin{aligned} -\check{\nabla}^2\psi'&=0 \quad \text{in } S^+=\{\mathbf{y}\,|\,z(y_1,y_2)\le y_3\le D\},\\ -\check{\nabla}^2\psi&=0 \quad \text{in } S^-=\{\mathbf{y}\,|\,-D\le y_3\le z(y_1,y_2)\}. \end{aligned} \tag{2.5}$$

Equation (2.1) provides two equivalent conditions across the interface [17]:

(2.6a) $$[\mathbf{H}\cdot\mathbf{T}]=0$$

and

(2.6b) $$\psi'-\psi=0 \quad \text{at } y_3=z(y_1,y_2)$$

where $\mathbf{T}$ is any vector tangent to z and, in this section, $[\,\cdot\,]$ represents the jump across the interface, the value above less the value below.

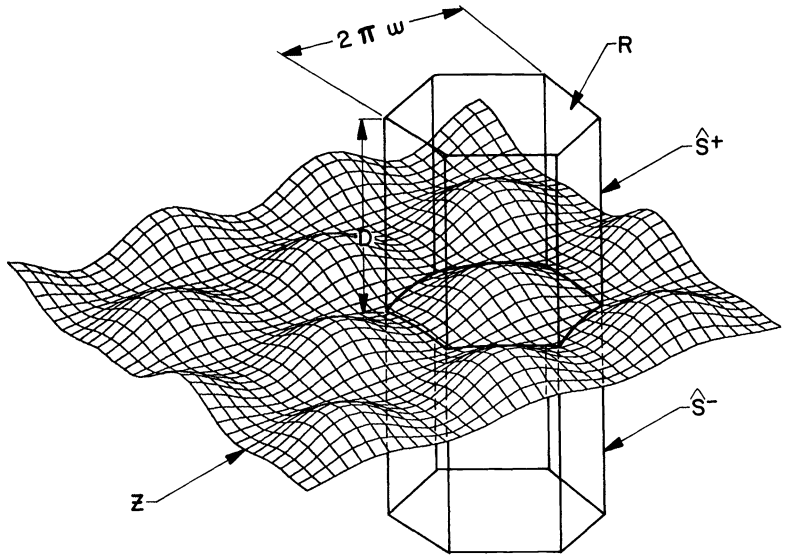


FIG 1. *The interface z when a hexagonal surface relief pattern has been established. Also, the sets associated with this pattern in the physical* $(y_1,y_2,y_3)^T$ *space.*

From (2.2) we get the condition

(2.7) $$[\mathbf{B}\cdot\mathbf{N}]=0$$

for $\mathbf{N}$ a vector normal to z. Specifically, we shall use

$$\mathbf{N}=(-z_1,-z_2,1)^T$$

where the function's subscripts represent partial derivatives with respect to y_i, $i=1,2,3$.

We next consider the stress tensor [1]

$$\sigma=-\left(p^*+\tfrac{1}{2}\mu_0\mathbf{H}^T\mathbf{H}\right)I+\mathbf{H}\mathbf{B}^T$$

where p^* is an effective pressure, to be eliminated presently, and I is the identity matrix. Any change in stress across the interface, $[\sigma\mathbf{N}]$, must be balanced by the surface tension, $\tau s\mathbf{N}$ [18], where

$$s=\frac{z_{11}(1+z_2^2)+z_{22}(1+z_1^2)-2z_1z_2z_{12}}{(1+z_1^2+z_2^2)^{3/2}}$$

and τ is the coefficient of surface tension.

Equation (2.6a) allows us to write

$$[\mathbf{H}]=-(\psi_3'-\psi_3)\mathbf{N}.$$

Thus $[\sigma \mathbf{N}]$ can be expressed as follows:

$$\begin{aligned}[\sigma \mathbf{N}] &= -\left[p^* + \tfrac{1}{2}\mu\mu_0 \mathbf{H}^T\mathbf{H}\right]\mathbf{N} + [\mathbf{H}\mathbf{B}^T\mathbf{N}] \\ &= -\left[p^* + \tfrac{1}{2}\mu\mu_0 \mathbf{H}^T\mathbf{H}\right]\mathbf{N} + [\mathbf{H}]\mathbf{B}^T\mathbf{N} \\ &= \left\{-\left[p^* + \tfrac{1}{2}\mu\mu_0 \mathbf{H}^T\mathbf{H}\right] - \mathbf{B}^T\mathbf{N}(\psi_3' - \psi_3)\right\}\mathbf{N}.\end{aligned}$$

Therefore the equality of the stress difference to the surface tension is equivalent to the equation

$$-\left[p^* + \tfrac{1}{2}\tilde{\mu}\mu_0 \mathbf{H}^T\mathbf{H}\right] - (\psi_3' - \psi_3)\mathbf{B}^T\mathbf{N} - \tau s = 0. \tag{2.8}$$

Note that this has reduced the vector equation involving the stress tensor to a single scalar equation.

We now consider the total forces, i.e., the body forces—gravity and the divergence of the stress tensor. With the following identity,

$$\begin{aligned}\check{\nabla} \cdot \mathbf{H}\mathbf{B}^T &= \mathbf{H}(\check{\nabla} \cdot \mathbf{B}) + (\mathbf{B} \cdot \check{\nabla})\mathbf{H} \\ &= \mathbf{H}(\check{\nabla} \cdot \mathbf{B}) + (\check{\nabla} \times \mathbf{H}) \times \mathbf{B} + \frac{\tilde{\mu}\mu_0}{2}\check{\nabla}(\mathbf{H}^T\mathbf{H}) \\ &= \frac{\tilde{\mu}\mu_0}{2}\check{\nabla}(\mathbf{H}^T\mathbf{H}),\end{aligned}$$

setting the sum of the forces equal to zero yields

$$\check{\nabla} \cdot \sigma + \rho g \check{\nabla} y_3 = -\check{\nabla}\left(p^* - \tfrac{1}{2}\mu_0(\tilde{\mu} - 1)\mathbf{H}^T\mathbf{H} - \rho g y_3\right) = 0$$

or

$$-\left(p^* - \tfrac{1}{2}\mu_0(\tilde{\mu} - 1)\mathbf{H}^T\mathbf{H} - \rho g y_3\right) = \text{constant}$$

where ρ is the fluid density and g is the acceleration due to gravity. Using the expression above to eliminate $[p^*]$ from (2.8) we get

$$-\Delta\rho g z - \tfrac{1}{2}[\mathbf{B}^T\mathbf{H}] - (\psi_3' - \psi_3)\mathbf{B}^T\mathbf{N} = -\tau s + \text{constant} \tag{2.9}$$

where $\Delta\rho$ is the change in density across the interface such that $\Delta\rho$ is positive. We assume when no magnetic field is present the planar interface is stable. No change in these equations would occur if the ferrofluid were less dense and gravity acted in the positive y_3 direction.

To put the problem in its final form we must include the bifurcation parameter H, the magnetic field strength at the lower boundary of the ferrofluid when the interface is a horizontal plane surface. This is accomplished by subtracting a trivial solution

$$z \equiv 0 \quad \text{(a planar surface)}, \qquad \psi' = \mu H y_3, \qquad \psi = H y_3$$

and since the planar surface was set at zero the constant is now forced to be $-(\mu\mu_0/\tau)((\mu - 1)/2)H^2$. Only a moderate amount of complexity would be added if a constant horizontal field were added to the trivial solution for linear stability of that case (see [10]). If we now replace $\mathbf{H}$ and $\mathbf{B}$ by the appropriate gradients, ψ' by

$\psi'+\mu Hy_3$, and ψ by $\psi+Hy_3$, we get

$$
\begin{aligned}
(2.10)\quad & -\check{\nabla}^2\psi'=0 \quad \text{in } S^+,\\
& -\check{\nabla}^2\psi=0 \quad \text{in } S^-,\\
& -\frac{\nabla\rho g}{\tau}z-\frac{\mu_0}{2\tau}\left(\left|\check{\nabla}(\psi'+\mu Hy_3)\right|^2-\mu\left|\check{\nabla}(\psi+\mu Hy_3)\right|^2\right)\\
& \qquad +\frac{\mu_0}{\tau}(\psi_3'+\mu H)\check{\nabla}(\psi'+\mu Hy_3)\cdot\mathbf{N}-\frac{\mu\mu_0}{\tau}(\psi_3+H)\check{\nabla}(\psi+Hy_3)\cdot\mathbf{N}\\
& \qquad -\frac{\mu\mu_0}{\tau}\frac{(\mu-1)}{2}H^2+s=0 \quad \text{at } y_3=z(y_1,y_2),\\
& \psi'-\psi+(\mu-1)Hz=0 \quad \text{at } y_3=z(y_1,y_2),\\
& \check{\nabla}(\psi'-\mu\psi)\cdot\mathbf{N}=0 \quad \text{at } y_3=z(y_1,y_2)
\end{aligned}
$$

in place of (2.5), (2.6b), (2.7) and (2.9).

We also now restrict the range of values y_1 and y_2 and impose boundary conditions. Experimental results produce two different surface patterns: squares and hexagons. To study one or the other of these we restrict (y_1,y_2) to a set R, where R is either a square or a hexagon in $\mathbb{R}^2$. To study the two-dimensional problem of rolls, let R be a rectangle with all function values independent of the y_2 coordinate.

Using an appropriate region R, we replace S^+ by $\check{S}^+$ and S^- by $\check{S}^-$ where

$$
\begin{aligned}
\check{S}^+&=\{\mathbf{y}|(y_1,y_2)\in R\}\cap S^+,\\
\check{S}^-&=\{\mathbf{y}|(y_1,y_2)\in R\}\cap S^-.
\end{aligned}
$$

To reflect the anticipated periodicity of the solution, we require that

$$
\begin{aligned}
(2.11)\quad & \frac{\partial z}{\partial n}=0 \quad \text{on } \partial R,\\
& \frac{\partial\psi'}{\partial n}=0 \quad \text{for } \mathbf{y}\in S^+ \text{ and } (y_1,y_2)\in\partial R,\\
& \frac{\partial\psi}{\partial n}=0 \quad \text{for } \mathbf{y}\in S^- \text{ and } (y_1,y_2)\in\partial R
\end{aligned}
$$

where $\frac{\partial}{\partial n}$ is the normal derivative.

The magnetic fluid must satisfy conditions at the upper and lower boundaries similar to those at the interface. Since ψ' and ψ are perturbations which decay toward the boundaries at $y_3=\pm D$,

$$
\begin{aligned}
(2.12)\quad & \psi_3'=0 \quad \text{at } y_3=D,\\
& \psi_3=0 \quad \text{at } y_3=-D.
\end{aligned}
$$

To ensure uniqueness of the solutions of equations (2.10) with boundary conditions (2.11) and (2.12) we also require that

$$
(2.13)\qquad \psi'|_{y_3=D}+\mu\psi|_{y_3=-D}=0.
$$

The five equations (2.10) and the boundary condition (2.11)–(2.13) are the statement of the problem which will be considered throughout the remainder of this article.

3. Mapping the free surface. In order to obtain a solution to equations (2.10) and the associated boundary conditions, we must determine the magnetostatic potentials ψ' and ψ and simultaneously the regions in which they are defined. In this situation the problem cannot easily be placed in any of the usual Banach spaces. To circumvent this

difficulty, we eliminate the free surface using the transformation

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dfrac{y_3 - z}{D - z} D \end{pmatrix} \quad \text{for } z < y_3 < D$$

and

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dfrac{y_3 - z}{D + z} D \end{pmatrix} \quad \text{for } -D < y_3 < z.$$

The techniques of tensor analysis [13], [19], [20] will allow us to map equations (2.10)–(2.13) to the $(x_1, x_2, x_3)^T$ space (Fig. 2).

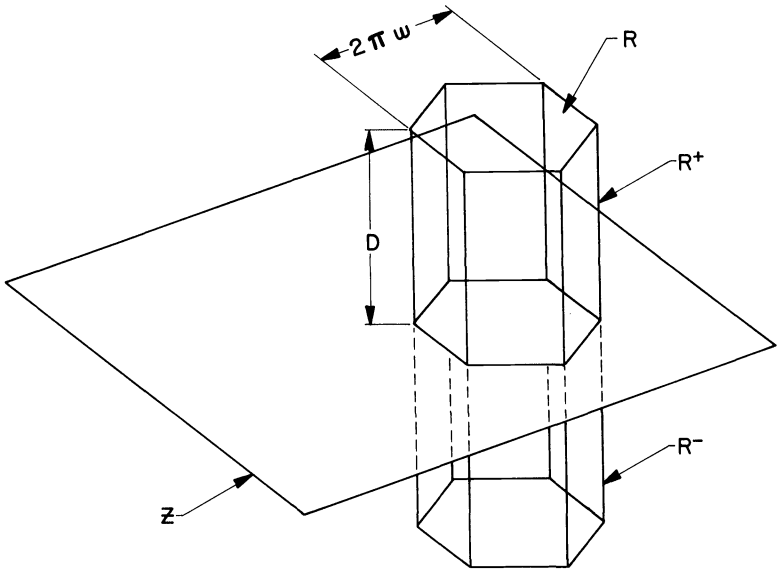


FIG 2. *The sets associated with the hexagonal surface relief pattern after the transformation has been made to the* (x_1, x_2, x_3) *space.*

Let

$$K_1 = \frac{D \mp y_3}{(D \mp z)^2} D \quad \text{and} \quad K_2 = \frac{D}{D \mp z},$$

where in this section, when a $\pm$ or $\mp$ symbol occurs, select the upper or lower symbol of the pair when the reference is in the volume above or below the interface, respectively. If we denote a function of (y_1, y_2, y_3) by $\check{f}$ and the same function in terms of (x_1, x_2, x_3) by f, then

$$\check{\nabla} \check{f} \to \begin{pmatrix} \nabla f - \nabla z K_1 f_3 \\ K_2 f_3 \end{pmatrix} = \begin{pmatrix} f_1 - z_1 K_1 f_3 \\ f_2 - z_2 K_1 f_3 \\ K_2 f_3 \end{pmatrix},$$

$$\check{\nabla}^2 \check{f} \to \nabla^2 f - \nabla^2 z K_1 f_3 - 2 K_1 \nabla z \cdot \nabla f_3 + |\nabla z|^2 K_1^2 f_{33} + K_2^2 f_{33} \mp 2 |\nabla z|^2 K_1 K_2 f_3 / D + |\nabla z|^2 f_3 K_1 K_2 / (z \mp D)$$

and

$$\mathbf{N}=\begin{pmatrix} -\partial z/\partial y_1 \\ -\partial z/\partial y_2 \\ 1 \end{pmatrix} \to \begin{pmatrix} -\nabla z \\ \left(1-|\nabla z|^2 K_1\right)/K_2 \end{pmatrix} = \begin{pmatrix} -z_1 \\ -z_2 \\ \left(1-|\nabla z|^2 K_1\right)/K_2 \end{pmatrix},$$

where $\nabla=(\partial/\partial x_1, \partial/\partial x_2)^T$ and the function's subscripts represent partial derivatives (one-sided if necessary) with respect to x_i, $i=1,2,3$.

Using the above quantities, (2.10) is equivalent to (3.1):

$$\begin{aligned}
(3.1)\qquad & -\nabla^2\phi' - K_2^2\phi'_{33} + \nabla^2 z K_1 \phi'_3 + 2K_1 \nabla z \cdot \nabla \phi'_3 - |\nabla z|^2 K_1^2 \phi'_{33} \\
& \quad + 2|\nabla z|^2 K_1 K_2 \phi'_3/D - |\nabla z|^2 \phi'_3 K_1 K_2/(z-D) = 0 \\
& \qquad\qquad \text{in } R^+ = \{\mathbf{x} \mid 0 < x_3 < D, (x_1, x_2) \in R\}, \\
& -\nabla^2\phi - K_2^2\phi_{33} + \nabla^2 z K_1 \phi_3 + 2K_1 \nabla z \cdot \nabla \phi'_3 - |\nabla z|^2 K_1^2 \phi_{33} \\
& \quad - 2|\nabla z|^2 K_1 K_2 \phi_3/D - |\nabla z|^2 \phi_3 K_1 K_2/(z+D) = 0 \\
& \qquad\qquad \text{in } R^- = \{\mathbf{x} \mid -D < x_3 < 0, (x_1, x_2) \in R\}, \\
& -s + \frac{\Delta\rho g}{\tau} z - \frac{\mu\mu_0}{\tau} H(\phi'_3 - \phi_3) + \frac{\mu\mu_0}{\tau} H \nabla z \cdot \nabla(\phi' - \phi) \\
& \quad + \frac{\mu_0}{2\tau}\left(|\nabla\phi'|^2 - \mu|\nabla\phi|^2 - \frac{D^2 - 2Dz}{(D-z)^2}\phi_3^{2\prime} + \mu \frac{D^2 - 2Dz}{(D-z)^2}\phi_3^2 \right) \\
& \qquad + \frac{\mu_0}{2\tau}|\nabla z|^2 \left(K_1^2 \phi_3^{2\prime} - \mu K_1^2 \phi_3^2\right) = 0 \quad \text{on } R, \\
& \phi' - \phi + (\mu - 1) H z = 0 \qquad \text{on } R, \\
& \phi'_3 - \mu\phi_3 - \nabla z \cdot \nabla(\phi' - \mu\phi) = 0 \quad \text{on } R.
\end{aligned}$$

and the boundary conditions remain unchanged:

$$\begin{aligned}
(3.2)\qquad & \frac{\partial z}{\partial n} = 0 \quad \text{on } \partial R, \\
& \frac{\partial \phi'}{\partial n} = 0 \quad \text{for } \mathbf{x} \in R^+ \text{ and } (x_1, x_2) \in \partial R, \\
& \frac{\partial \phi}{\partial n} = 0 \quad \text{for } \mathbf{x} \in R^- \text{ and } (x_1, x_2) \in \partial R, \\
& \phi'_3 = 0 \quad \text{for } x_3 = D, \\
& \phi_3 = 0 \quad \text{for } x_3 = -D, \\
& \phi'|_{x_3 = D} + \mu\phi|_{x_3 = -D} = 0
\end{aligned}$$

where ϕ' and ϕ are ψ' and ψ under the new coordinate system. This transformation leaves $(\phi', \phi, z)^T = \mathbf{0} = (0,0,0)^T$ as a solution to (3.1) and (3.2).

The problem can now be placed in a Hölder space setting. Let $\| \ \|_{K+\alpha}$ denote the usual norm on the Hölder space $C^{K+\alpha}(R)$ [4]. Similarly, let $\| \ \|^+_{K+\alpha}$ and $\| \ \|^-_{K+\alpha}$ denote norms on $C^{K+\alpha}(R^+)$ and $C^{K+\alpha}(R^-)$, respectively. Let

$$B_{K,l,\alpha} = C^{K+\alpha}(R^+) \times C^{K+\alpha}(R^-) \times C^{K+\alpha}(R) \times C^{l+\alpha}(R) \times C^{l+\alpha}(R)$$

be a Banach space whose norm on $\mathbf{f}=(f^1,f^2,f^3,f^4,f^5)$ is defined as

$$|\mathbf{f}|_{K,l,\alpha}=\|f^1\|^{+}_{K+\alpha}+\|f^2\|^{-}_{K+\alpha}+\|f^3\|_{K+\alpha}+\|f^4\|_{l+\alpha}+\|f^5\|_{l+\alpha}.$$

Let $B_{2+\alpha}$ be equal to the subspace of $B_{2,1,\alpha}$ such that

$$\frac{\partial f^1}{\partial n}=0 \quad \text{for } \mathbf{x}\in R^+ \text{ and } (x_1,x_2)\in\partial R,$$
$$\frac{\partial f^2}{\partial n}=0 \quad \text{for } \mathbf{x}\in R^- \text{ and } (x_1,x_2)\in\partial R,$$
$$\frac{\partial f^3}{\partial n}=0 \quad \text{on } \partial R,$$
$$f^1_3=0 \quad \text{at } x_3=D,$$
$$f^2_3=0 \quad \text{at } x_3=-D,$$
$$f^1\big|_{x_3=D}+\mu f^2\big|_{x_3=-D}=0,$$
$$f^4-\frac{4}{\omega}f^1_3\big|_{x_3=0}=0,$$
$$f^5-f^1_3\big|_{x_3=0}+f^2_3\big|_{x_3=0}=0,$$

with $|\cdot|_{2+\alpha}$ denoting the associated norm and ω the solution to (4.2). The above conditions define f^4 and f^5. These functions are used to normalize the linear operator defined in §4. Finally, let $B_\alpha=B_{0,0,\alpha}$ and $|\cdot|_\alpha$ be the associated norm.

Using the above definitions, the following operators continuously map $B_{2+\alpha}$ to B_α and will be used extensively in the remaining sections:

$$L_0\boldsymbol{\phi}=\begin{pmatrix} -\nabla^2\phi'-\phi'_{33} \\ -\nabla^2\phi-\phi_{33} \\ \dfrac{\Delta\rho g}{\tau}z-\nabla^2 z-\dfrac{\mu\mu_0}{\tau}H_C\delta \\ \phi'-\phi+(\mu-1)H_C z \\ \mu\delta-\dfrac{\mu-1}{4}\sqrt{\dfrac{\Delta\rho g}{\tau}}\,\gamma \end{pmatrix},$$

$$L_1\boldsymbol{\phi}=\begin{pmatrix} 0 \\ 0 \\ -\dfrac{\mu\mu_0}{\tau}\delta \\ (\mu-1)z \\ 0 \end{pmatrix},$$

$$N_1(\boldsymbol{\phi},\boldsymbol{\phi})=\begin{pmatrix} \nabla^2 z\phi'_3(1-x_3/D)+2\nabla z\cdot\nabla\phi'_3(1-x_3/D)-2\phi'_{33}z/D \\ \nabla^2 z\phi_3(1+x_3/D)+2\nabla z\cdot\nabla\phi_3(1+x_3/D)+2\phi_{33}z/D \\ \dfrac{\mu_0}{2\tau}\left(|\nabla\phi'|^2-\mu|\nabla\phi|^2-\phi_3^{2\prime}+\mu\phi_3^2\right)+\dfrac{\mu\mu_0}{\tau}H_C\nabla z\cdot\nabla(\phi'-\phi) \\ -\nabla z\cdot\nabla(\phi'-\mu\phi) \end{pmatrix},$$

$$N_2(\phi,\phi)=\begin{pmatrix} 0 \\ 0 \\ \frac{\mu\mu_0}{\tau}\nabla z\cdot\nabla(\phi'-\phi) \\ 0 \\ 0 \end{pmatrix},$$

$$N_3(\phi,\phi,\phi)=\begin{pmatrix} -|\nabla z|^2\phi'_{33}(1-x_3/D)-3\phi'_{33}z^2/D^2-(\nabla^2 z)\phi'_3 z/D(x_3/D-1) \\ -2\nabla z\cdot\nabla\phi'_3 z/D(x_3/D-1)+|\nabla z|^2\phi'_3/D(1-x_3/D) \\ -|\nabla z|^2\phi_{33}(1+x_3/D)-3\phi_{33}z^2/D^2-(\nabla^2 z)\phi_3 z/D(x_3/D+1) \\ -2\nabla z\cdot\nabla\phi_3 z/D(x_3/D+1)-|\nabla z|^2\phi_3/D(1+x_3/D) \\ -z_{11}z_2^2-z_{22}z_1^2+2z_1z_2z_{12}+3|\nabla z|^2\nabla^2 z \\ 0 \\ 0 \end{pmatrix},$$

where

$$\gamma=4\sqrt{\frac{\tau}{\Delta\rho g}}\,\phi'_3\Big|_{x_3=\theta},$$

$$\delta=\phi'_3|_{x_3=0}-\phi_3|_{x_3=0},$$

$$\phi=(\phi',\phi,z,\gamma,\delta)^T$$

and H_c is a constant to be determined later. These operators allow us to write (3.1) as

$$F(H,\phi)=L_0\phi+(H-H_c)L_1\phi+N_1(\phi,\phi)+(H-H_c)N_2(\phi,\phi)+N_3(\phi,\phi,\phi)+\mathrm{Rem}(H,\phi)=\mathbf{0}$$

where Rem contains all higher order terms of F. The linear boundary conditions (3.2) are contained in the domain space $B_{2+\alpha}$.

4. Linearized eigenvalue problem and associated adjoint eigenvalue problem. We will now determine the first possible value of H for which the trivial solution could lose stability, that is, the minimum positive eigenvalue of the Fréchet derivative of F. The associated eigenvector will be used later as the first term in a series expansion of a bifurcating branch of solutions. Also associated with the Fréchet derivative at the minimum eigenvalue is an adjoint operator with the same eigenvalue. The eigenvector of the adjoint operator will allow us to define a projection and ultimately use the implicit function theorem to prove existence and stability results.

The Fréchet derivative $F_\phi(H_c, \mathbf{0})$ with respect to ϕ at $(H_c, \mathbf{0})$ is L_0 where H_c is the minimum eigenvalue (the critical magnetic field strength) for which we are searching. Thus we must find a nontrivial vector ϕ_0 and the minimum positive value, H_c, such that $L_0\phi_0 = \mathbf{0}$.

Since ϕ and ϕ' must satisfy Laplace's equation and the normal derivative must be zero in the horizontal directions, x_1 and x_2, we look for solutions of the form

$$\phi' = \sum_{j=1}^{\infty} \left(A_j e^{-j\omega x_3} + C_j e^{j\omega x_3}\right) \mathrm{Tr}_j(x_1, x_2),$$

$$\phi = \sum_{j=1}^{\infty} \left(B_j e^{j\omega x_3} + D_j e^{-j\omega x_3}\right) \mathrm{Tr}_j(x_1, x_2)$$

where A_j, B_j, C_j and D_j are real constants and Tr_j satisfies, for all j, the following:

$$\nabla^2 \mathrm{Tr}_j = -\omega^2 j^2 \mathrm{Tr}_j \quad \text{in } R, \tag{4.1}$$

$$\frac{\partial \mathrm{Tr}_j}{\partial n} = 0 \quad \text{on } \partial R.$$

With the appropriate choice of Tr_j's and R the results can apply to either rolls, rectangles or hexagons. Hence the results in the following sections apply equally to all three cases.

If we use the above to solve $L_0\phi_0 = \mathbf{0}$ in $B_{2+\alpha}$ we get

$$\phi_0 = \begin{bmatrix} -\mu\left(e^{-\omega x_3} + e^{-\omega(2D - x_3)}\right)\mathrm{Tr} \\ \left(e^{\omega x_3} + e^{-\omega(2D + x_3)}\right)\mathrm{Tr} \\ \dfrac{\mu+1}{\mu-1}\dfrac{1}{H}\left(1 + e^{-2\omega D}\right)\mathrm{Tr} \\ 4\mu\left(1 - e^{-2\omega D}\right)\mathrm{Tr} \\ (\mu-1)\omega\left(1 - e^{-2\omega D}\right)\mathrm{Tr} \end{bmatrix}$$

where $\mathrm{Tr} = \mathrm{Tr}_1$ and

$$H_c^2 = \frac{\mu+1}{(\mu-1)^2}\frac{\tau}{\mu\mu_0}\left(\frac{\Delta\rho g}{\tau} + \omega^2\right)\frac{\coth \omega D}{\omega}$$

is the associated eigenvalue. To obtain the minimum positive eigenvalue we minimize H_c^2 with respect to ω. This minimum occurs when

$$f(\omega, D) = \left(\frac{\Delta\rho g}{\tau} + \omega^2\right)\omega D - \left(\frac{\Delta\rho g}{\tau} - \omega^2\right)\cosh 2\omega D \tag{4.2}$$

is equal to zero. For $\omega_u = \sqrt{\Delta\rho g/\tau}$, $f(\omega_u, D)$ is positive. When $\omega_L = \sqrt{D/(D+1)\,\Delta\rho g/\tau}$,

$$f(\omega_L, D) = \frac{\Delta\rho g}{\tau}\frac{1}{D+1}\left[2\omega D^2 + \omega D - \cosh 2\omega D\right]$$

is less than zero for large D (since the hyperbolic cosine grows more rapidly than D^2). Hence,

$$\sqrt{\frac{D}{D+1}\frac{\Delta\rho g}{\tau}} < \omega < \sqrt{\frac{\Delta\rho g}{\tau}}$$

and

$$\frac{(\mu-1)^2}{2(\mu+1)}\frac{\mu\mu_0 H_c^2}{\sqrt{\Delta\rho g\tau}}<\sqrt{\frac{D+1}{D}}\cosh D\sqrt{\frac{\Delta\rho g}{\tau}}$$

for large D.

We note that for D infinite, this forces the same value of H_c as obtained by Cowley and Rosensweig [1] in their linear stability analysis.

We next define a linear functional on B_α and $B_{2+\alpha}$:

$$\langle \phi, \mathbf{u}\rangle = \int_{R^+} \phi' u^1\, dv + \int_{R^-} \phi u^2\, dv + \int_R \left(zu^3+\gamma u^4+\delta u^5\right) d\sigma$$

for a fixed $\mathbf{u}=(u^1,u^2,u^3,u^4,u^5)^T$. Let $\mathbf{u}^*$ satisfy $\langle L_0\phi, \mathbf{u}^*\rangle = 0$ for all ϕ in $B_{2+\alpha}$. Solving for $\mathbf{u}^*$ using integration by parts, we obtain

$$\mathbf{u}^* = \frac{1}{k}\begin{bmatrix} -\left(e^{-\omega x_3}+e^{-\omega(2D-x_3)}\right)\mathrm{Tr} \\ \left(e^{\omega x_3}+e^{-\omega(2D+x_3)}\right)\mathrm{Tr} \\ -\dfrac{\mu+1}{\mu-1}\dfrac{\tau}{\mu\mu_0}\dfrac{1}{H_c}\left(1+e^{-2\omega D}\right)\mathrm{Tr} \\ \omega\left(1-e^{-2\omega D}\right)\mathrm{Tr} \\ -\dfrac{2}{\mu-1}\left(1+e^{-2\omega D}\right)\mathrm{Tr} \end{bmatrix}.$$

With the additional condition $\langle \phi_0, \mathbf{u}^*\rangle = 1$ we find that

$$k=\langle \phi_0, k\mathbf{u}^*\rangle = \left\{-\left(2\omega-\frac{\mu+1}{2\omega}\right)\left(1-e^{-4\omega D}\right)-\frac{(\mu+1)\omega}{\Delta\rho g/\tau+\omega^2}\left(1-e^{-4\omega D}\right)\right.$$
$$\left.+4\mu\omega\left(1-e^{-2\omega D}\right)^2+2D(\mu+1)e^{-2\omega D}\right\}\int_R \mathrm{Tr}^2.$$

For D sufficiently large, k is positive. We now restrict D to this range and define the continuous linear functional $[\,\cdot\,]$ by

$$[\,\cdot\,] = \langle \cdot, \mathbf{u}^*\rangle.$$

5. Preliminary lemmas. Some standard results and several lemmas are presented in this section. These will be used in the proofs of the stability and existence of the bifurcating branch.

Operators carrying a vector subscript are Fréchet derivatives of the operator with respect to that vector.

The proofs of the existence and stability theorems rely heavily on the implicit function theorem. We shall use the following form [4].

THEOREM 1 (implicit function theorem (IFT)). *If* $T:\mathbb{R}\times B_1 \to B_2$ *is continuously differentiable in* $(\lambda,\mathbf{x})$ *in a neighborhood of* $(0,\mathbf{0})$, $T(\lambda,\mathbf{0})=\mathbf{0}$, *and* $T_{\mathbf{x}}(0,\mathbf{0})$ *is continuously invertible from* B_2 *to* B_1, *then there exists for small* $|\lambda|$ *a function* $\mathbf{x}(\lambda): |\mathbb{R} \to B_1$ *such that*

a) $\mathbf{x}(0)=\mathbf{0}$,

b) $\mathbf{x}$ *has a continuous derivative with respect to* λ,

c) $T(\lambda,\mathbf{x}(\lambda))=\mathbf{0}$.

When we use IFT, λ and $\mathbf{x}$ are expressed as power series expansions. The convergence of these expansions follows from the analyticity of T [4]:

THEOREM 2 (analyticity). *If the operator T in* IFT *is analytic in some neighborhood of* $(0,\mathbf{0})$, *then* $\mathbf{x}$ *is also analytic in* λ *in some neighborhood of* 0.

We cannot apply the IFT to the function F since we already know that $F_\phi(0,\mathbf{0})$ is not invertible ($h=H-H_c=0$ is an eigenvalue). We must reformulate the problem to project along this eigenfunction using $[\cdot]$. The reformulation allows us to obtain a generalized inverse to $F_\phi(0,\mathbf{0})$. The continuity of the inverse is the subject of Lemmas 2 to 4, and is based on a priori Schauder estimates given below.

THEOREM 3 (a priori Schauder estimates). 1) *Let*

$$\nabla^2 u-\omega^2 u=f \quad \text{on } R$$

and

$$\frac{\partial u}{\partial n}=0 \quad \text{on } \partial R$$

where $u\in C^{2+\alpha}(R), f\in C^{\alpha}(R)$ *and* $\omega^2>0$. *Then*

$$\|u\|_{2+\alpha}<\text{const}\left(\|f\|_\alpha\right).$$

2) *Let*

$$\nabla^2 u=f \quad \text{in } R^+,$$
$$u=g \quad \text{when } x_3=0,$$
$$\frac{\partial u}{\partial x_3}=h \quad \text{when } x_3=-D,$$
$$\frac{\partial u}{\partial n}=\frac{\partial f}{\partial n}=0 \quad \text{when } (x_1,x_2)\in\partial R \text{ and } 0\le x_3\le D,$$
$$\frac{\partial g}{\partial n}=\frac{\partial h}{\partial n}=0 \quad \text{on } \partial R,$$

where $u\in C^{2+\alpha}(R^+), f\in C^{\alpha}(R^+)$, $g\in C^{2+\alpha}(R)$, *and* $h\in C^{1+\alpha}(R)$. *Then*

$$\|u\|^{+}_{2+\alpha}\le\text{const}\left(\|f\|^{+}_{\alpha}+\|g\|_{2+\alpha}+\|h\|_{1+\alpha}\right).$$

Proof. The above estimates follow from using the usual interior estimates for elliptic problems, the boundary estimates for Dirichlet problems, the boundary estimates for the oblique derivative problem and the periodicity of our domains. (See [21, Thms. 6.2, 6.6 and 6.30].)

Some needed facts are stated in Lemma 1. Condition i) implies that H_c is an algebraically simple eigenvalue [4].

LEMMA 1. *Let* L_0 *and* L_1 *be the operators defined in* §3, *and let* ϕ_0 *and* $[\cdot]$ *be as given in* §4. *Then, for sufficiently large* D, *the following hold*:

(i) $[\phi_0]=1$,
(ii) $[L_1\phi_0]>0$,
(iii) $[L_0\phi]=0$ *for all* $\phi\in B_{2+\alpha}$.

The proof is straightforward and will not be given here.

LEMMA 2. *The operator* $A: B_{2+\alpha} \to B$ *defined by*

$$A\boldsymbol{\phi}=A\begin{bmatrix}\phi'\\ \phi\\ z\\ \gamma\\ \delta\end{bmatrix}=\begin{bmatrix}-\nabla^2\phi'-\phi'_{33}\\ -\nabla^2\phi-\phi_{33}\\ -\nabla^2 z+\dfrac{\Delta\rho g}{\tau}z\\ \phi'-\phi\\ \phi'_3-\mu\phi_3\end{bmatrix}=\begin{bmatrix}f^1\\ f^2\\ f^3\\ f^4\\ f^5\end{bmatrix}=\mathbf{f}$$

has a continuous inverse.

Proof. If we make the following change of coordinates

$$x'_1=x_1, \qquad x'_2=x_2,$$
$$x'_3=-x_3+D \quad \text{in } R^+, \qquad x'_3=x_3+D \quad \text{in } R^-,$$

then the following equations must be satisfied.

$$-\nabla^2(\phi'+\mu\phi)=f_1+\mu f_2 \quad \text{in } R^+,$$
$$\phi'_3+\mu\phi_3=0 \quad \text{at } x_3=D,$$
$$\phi'+\mu\phi=f_5 \quad \text{at } x_3=0.$$

Using the a priori estimates, we get our first relation:

$$\|\phi'+\mu\phi\|_{2+\alpha}<\text{const}\left(\|f_1+\mu f_2\|_\alpha+\|f_5\|_{1+\alpha}\right).$$

Similarly, with the following change of coordinates:

$$x'_1=x_1, \qquad x'_2=x_2,$$
$$x'_3=x_3 \quad \text{in } R^+, \qquad x'_3=-x_3 \quad \text{in } R^-,$$

we obtain the equations

$$\nabla^2(\phi'-\phi)=f_1-f_2 \quad \text{in } R^+,$$
$$\phi'_3-\phi_3=0 \quad \text{at } x_3=D,$$
$$\phi'-\phi=f_4 \quad \text{at } x_3=0.$$

Then the a priori estimates result in

$$\|\phi'-\phi\|_{2+\alpha}\leq\text{const}\left(\|f_1-f_2\|_\alpha+\|f_4\|_{2+\alpha}\right).$$

Finally, we have immediately from Theorem 3

$$\|z\|_{2+\alpha}<\text{const}\|f_3\|_\alpha.$$

These inequalities can be combined with the backwards triangle inequality to show that

$$|\boldsymbol{\phi}|_{2+\alpha}<\text{const}|A\boldsymbol{\phi}|_\alpha.$$

LEMMA 3. *If* $(L_0-\lambda I)\boldsymbol{\phi}=\mathbf{0}$ *has no nontrivial solutions, then* $L_0-\lambda I$ *has a continuous inverse.*

Proof. Let

$$M\boldsymbol{\phi}=\begin{pmatrix}0\\ 0\\ \dfrac{-\mu\mu_0}{\tau}\delta\\ 0\\ (\mu-1)H_c z\end{pmatrix}$$

where $\boldsymbol{\phi}=(\phi',\phi,z,\gamma,\delta)^T$. Then

$$L_0-\lambda I=A+(M-\lambda I).$$

Since A has a continuous inverse and $M-\lambda I$ is bounded from B_α to B_α, then $A^{-1}(M-\lambda I)$ is bounded from B_α to $B_{2+\alpha}$. If we apply Adams [22, Thm. 1.3.1], we see that $B_{2+\alpha}$ is embedded compactly in B_α since the domains are finite. Hence the operator $A^{-1}(M-\lambda I)$ is compact from $B_{2+\alpha}$ to $B_{2+\alpha}$.

Let $f\in B_\alpha$ and consider the equation:

$$(L_0-\lambda I)\boldsymbol{\phi}=\mathbf{f}$$

where $\boldsymbol{\phi}\in B_{2+\alpha}$ is unknown. The hypothesis that the equation $(L_0-\lambda I)\boldsymbol{\phi}=\mathbf{0}$ has no nontrivial solutions allows us to apply the Fredholm alternative [4] and obtain the existence of a unique $\boldsymbol{\phi}$ such that

$$(L_0-\lambda I)\boldsymbol{\phi}=\mathbf{f}$$

or, equivalently,

$$\left(I-A^{-1}(M-\lambda I)\boldsymbol{\phi}=A^{-1}\mathbf{f}\right)$$

which satisfies the following inequality:

$$|\boldsymbol{\phi}|_{2+\alpha}\leq \text{const}|A^{-1}\mathbf{f}|_{2+\alpha}.$$

Then, by Lemma 2, we obtain

$$|\boldsymbol{\phi}|_{2+\alpha}\leq \text{const}|\mathbf{f}|_\alpha=\text{const}|(L_0-\lambda I)\boldsymbol{\phi}|_\alpha,$$

which is what we were to prove.

We shall now prove the existence and continuity of the generalized inverse of L_0.

LEMMA 4. *Let*

$$P\mathbf{u}=[\mathbf{u}]\boldsymbol{\phi}_0 \quad \textit{and} \quad Q=I-P.$$

The operators P and Q are bounded projections from $B_{2+\alpha}$ to $B_{2+\alpha}$, and there exists a bounded linear operator K: $B_\alpha\to B_{2+\alpha}$ such that $KL_0=Q$.

Proof. Consider the equation

$$L_0\boldsymbol{\phi}=\mathbf{f}$$

or equivalently

$$\left(I+\lambda(L_0-\lambda I)^{-1}\right)\boldsymbol{\phi}=(L_0-\lambda I)^{-1}\mathbf{f} \tag{5.1}$$

where λ is not in the spectrum of L_0. Equation (5.1) has a nontrivial solution ($\boldsymbol{\phi}=\boldsymbol{\phi}_0$ for $\mathbf{f}=0$) which is not in the Banach space

$$\check{B}=\{\boldsymbol{\phi}\in B_{2+\alpha}|[\boldsymbol{\phi}]=0\}.$$

In fact, L_0 is one-to-one from $\check{B}$ since H_c was a simple eigenvalue. In Lemma 3 we showed that $(L_0-\lambda I)^{-1}$ was bounded from B_α to $B_{2+\alpha}$. Thus $(L_0-\lambda I)^{-1}$ must be compact from $\check{B}$ to $B_{2+\alpha}$ since the domains are bounded [22]. By applying the Fredholm alternative we know that $I+\lambda(L_0-\lambda I)^{-1}$ has a bounded inverse. Let $\check{K}$ denote the composition of $(I+\lambda(L_0-\lambda I)^{-1})^{-1}$ with $(L_0-\lambda I)^{-1}$. Then for $\phi\in\check{B}$, $\check{K}$ satisfies

$$\check{K}L_0\boldsymbol{\phi}=\boldsymbol{\phi}.$$

Using Lemma 1 it can easily be shown that P is a bounded projection. Since P and I are bounded projections, Q also has this property. Also note that Q maps $B_{2+\alpha}$ to $\check{B}$. If

we define $K=\check{K}Q$, then K is the bounded generalized inverse of L_0 and

$$KL_0=\check{K}QL_0=\check{K}L_0Q=Q^2=Q.$$

The proofs of Lemmas 3 and 4 required that the functions' domains be of finite extent. If D is allowed to be infinite, other techniques must be used to obtain the existence of the generalized inverse of L_0.

6. Existence. In this section, we prove the existence and analyticity of a bifurcating branch of solutions emanating from the trivial branch at H_c.

THEOREM 4 (existence). *There exists a nontrivial branch of solutions to* $F(H,\phi)=\mathbf{0}$ *which can be expressed parametrically as* $(H(\varepsilon),\phi(\varepsilon))$ *where* $H(0)=H_c$ *and* $\phi(0)=\phi_0$. *The series representations of* H *and* ϕ *are convergent for small* $|\varepsilon|$.

Proof. This proof is based on an application of IFT to the Lyapunov–Schmidt equations. Recall that

$$F(H,\phi)=L_0\phi+(H-H_c)L_1\phi+N_1(\phi,\phi)+N(H,\phi)$$

where

$$N(H,\phi)=(H-H_c)N_2(\phi,\phi)+N_3(\phi,\phi,\phi)+\mathrm{Rem}(H,\phi).$$

If we set $H-H_c=\varepsilon\eta$ and $\phi=\varepsilon\mathbf{w}$, where $\mathbf{w}$ is normalized so that $[\mathbf{w}]=1$, then N is of order ε^3 and $P\mathbf{w}=\phi_0$. We now write ϕ as $\varepsilon(\phi_0+\xi)$ where $\xi=Q\mathbf{w}$ and $[\xi]=0$. Substituting this into F and applying $[\,\cdot\,]$ and K, we get

$$\begin{aligned}&[\eta L_1(\phi_0+\xi)+N_1(\phi_0+\xi,\phi_0+\xi)+\varepsilon^{-2}N(H,\phi)]=0,\\ &\xi+\varepsilon K\{\eta L_1(\phi_0+\xi)+N_1(\phi_0+\xi,\phi_0+\xi)+\varepsilon^{-1}N(H,\phi)\}=0.\end{aligned}\tag{6.1}$$

Equations (6.1) are the Lyapunov–Schmidt equations. Individually they contain the same information as projecting by P and Q, respectively. Thus, they are equivalent to the original equation. If we define

$$T:\mathbb{R}\times(\mathbb{R}\times\check{B})\to\mathbb{R}\times B_{2+\alpha}$$

by

$$\begin{pmatrix}\varepsilon\\ \eta\\ \xi\end{pmatrix}\to\begin{pmatrix}[\eta L_1(\phi_0+\xi)+N_1(\phi_0+\xi,\phi_0+\xi)+\varepsilon^{-2}N(H,\phi)]\\ \xi+\varepsilon K\{\eta L_1(\phi_0+\xi)+N_1(\phi_0+\xi,\phi_0+\xi)+\varepsilon^{-1}N(H,\phi)\}\end{pmatrix},$$

one solution to $T(\varepsilon,\eta,\xi)\equiv\mathbf{0}$ is

$$\varepsilon=0,\quad \xi=\mathbf{0},\quad \eta_0=-\frac{[N_1(\phi_0,\phi_0)]}{[L_1\phi_0]}.$$

Since $[L_1\phi_0]>0$ by Lemma 1, η_0 is well defined. In order to apply IFT we note that

$$T_{(\eta,\xi)}(0,\eta_0,\mathbf{0})=\begin{pmatrix}\eta_0[L_1\,\cdot\,]+[N_1(\cdot\,,\phi_0)+N_1(\phi_0,\,\cdot\,)] & [L_1\phi_0]\\ I & 0\end{pmatrix},$$

which is invertible since $[L_1\phi_0]>0$. The existence of a branch of solutions is then a result of the IFT.

The convergent series representations for H and ϕ are obtained from the analyticity of F and Theorem 2.

7. Formal stability theory. In this section we prove our second major theorem. The stability is "formal" since we assume the physical stability of the system is equivalent to the linear stability. Proving this equivalence is a topic for future consideration. We

should also add that the stability given here is along the nontrivial solution branch whose existence was shown in §6, and not the stability of the trivial solutions, which is what most previous authors have considered.

THEOREM 5 (stability). *Consider the following expansion of H and ϕ in terms of ε:*

$$H=H_c+\varepsilon H_0+\varepsilon^2 H_1+\cdots,$$
$$\phi=\varepsilon\phi_0+\varepsilon^2\phi_1+\cdots.$$

Then $H_0=-[N_1(\phi_0,\phi_0)]/[L_1\phi_0]$, and the branch is stable if $H_0>0$ or unstable if $H_0<0$. In the case that $H_0=0$,

$$H_1=\frac{[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)+N_3(\phi_0,\phi_0,\phi_0)]}{[L_1\phi_0]},$$

and the branch is stable or unstable if H_1 is greater or less than zero, respectively.

We note that these results imply supercritical branches are stable and subcritical branches are unstable.

Proof. The convergence of the expansions of H and ϕ for small $|\varepsilon|$ is a result of the existence theorem. Expressions for ϕ_i and H_i, $i=0,1,2,\cdots$ follow from the equation $F(H,\phi)=\mathbf{0}$:

$$\begin{aligned}(7.1)\qquad F(H,\phi)=&\varepsilon L_0\phi_0+\varepsilon^2(L_0\phi_1+H_0L_1\phi_0+N_1(\phi_0,\phi_0))\\&+\varepsilon^3(L_0\phi_2+H_0L_1\phi_1+H_1L_1\phi_0\\&\quad+N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)\\&\quad+H_0N_2(\phi_0,\phi_0)+N_3(\phi_0,\phi_0,\phi_0))\\&+\text{higher order terms in }\varepsilon.\end{aligned}$$

Equating the coefficients of ε and ε^2 to zero gives the following equations:

$$L_0\phi_0=\mathbf{0}$$

and

$$L_0\phi_1+H_0L_1\phi_0+N_1(\phi_0,\phi_0)=\mathbf{0}.$$

The first equation has already been satisfied. Operating on the second equation by $[\cdot]$, we obtain the desired expression:

$$H_0=-\frac{[N_1(\phi_0,\phi_0)]}{[L_1\phi_0]}.$$

(Recall that $[L_0\phi_0]=0$ and $[L_1\phi_0]>0$ by Lemma 1.) If $H_0=0$, applying the same technique to the coefficients of ε^2 and ε^3 in (7.1) yields

$$(7.2)\qquad L_0\phi_1=-N_1(\phi_0,\phi_0)$$

and

$$H_1=-\frac{[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)+N_2(\phi_0,\phi_0,\phi_0)]}{[L_1\phi_0]}.$$

To study the stability of the solution (H,ϕ) of $F(H,\phi)=\mathbf{0}$, linear stability requires that we consider the equation

$$(7.3)\qquad F_\phi(H,\phi)\mathbf{u}=\sigma\mathbf{u}$$

where

$$\mathbf{u}=\boldsymbol{\phi}_0+\varepsilon\mathbf{u}_1+\varepsilon^2\mathbf{u}_2+\cdots,$$
$$\sigma=\varepsilon\sigma_1+\varepsilon^2\sigma_2+\cdots,$$
$$\boldsymbol{\phi}=\boldsymbol{\phi}_0+\varepsilon\boldsymbol{\phi}_1+\varepsilon^2\boldsymbol{\phi}_2+\cdots,$$
$$H=H_c+\varepsilon H_1+\varepsilon^2 H_2+\cdots$$

and $[\mathbf{u}]=1$. When $\sigma<0$ any local oscillations around $\boldsymbol{\phi}$ would decay or when $\sigma>0$ the oscillations would grow.

If we now substitute the expansions above into (7.3) we get

$$\begin{aligned}(7.4)\quad F_{\boldsymbol{\phi}}(H,\boldsymbol{\phi})\mathbf{u}-\sigma\mathbf{u}&=(L_0+(H-H_c)L_1)\mathbf{u}-\sigma\mathbf{u}\\&\quad+N_1(\boldsymbol{\phi},\mathbf{u})+N_1(\mathbf{u},\boldsymbol{\phi})\\&\quad+(H-H_c)(N_2(\boldsymbol{\phi},\mathbf{u})+N_2(\mathbf{u},\boldsymbol{\phi}))\\&\quad+N_3(\mathbf{u},\boldsymbol{\phi},\boldsymbol{\phi})+N_3(\boldsymbol{\phi},\mathbf{u},\boldsymbol{\phi})+N_3(\boldsymbol{\phi},\boldsymbol{\phi},\mathbf{u})\\&\quad+\text{higher order terms in }\boldsymbol{\phi}\\&=L_0\boldsymbol{\phi}_0+\varepsilon(L_0\mathbf{u}_1+H_0L_1\boldsymbol{\phi}_0+2N_1(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)-\sigma_1\boldsymbol{\phi}_0)\\&\quad+\varepsilon^2(L_0\mathbf{u}_2+H_0L_1\mathbf{u}_1+H_1L_1\boldsymbol{\phi}_0\\&\qquad+N_1(\boldsymbol{\phi}_0,\mathbf{u}_1)+N_1(\boldsymbol{\phi}_0,\boldsymbol{\phi}_1)+N_1(\boldsymbol{\phi}_1,\boldsymbol{\phi}_0)+N_1(\mathbf{u}_1\boldsymbol{\phi}_0)\\&\qquad+2H_0N_2(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)+3N_3(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)-\sigma_1\mathbf{u}_1-\sigma_2\boldsymbol{\phi}_0)\\&\quad+\text{higher order terms in }\varepsilon\\&=\mathbf{0}.\end{aligned}$$

Equating the coefficients of 1 and ε to zero we obtain

$$L_0\boldsymbol{\phi}_0=\mathbf{0}$$

and

$$L_0\mathbf{u}_1+H_0L_1\boldsymbol{\phi}_0+2N_1(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)-\sigma_1\boldsymbol{\phi}_0=0.$$

The first equation is already satisfied. Operating on the second by $[\cdot]$ and using the expression for H_0 determined above, we get

$$\sigma_1=-H_0[L_1\boldsymbol{\phi}_0].$$

Since $[L_1\boldsymbol{\phi}_0]>0$ the sign of σ_1 is opposite that of H_0. Hence the stability associated with nonzero H_0 is established.

In the case of zero H_0,σ_1 is also zero. By equating the coefficients of ε and ε^2 to zero in (7.4) with $H_0=\sigma_1=0$, we get

$$(7.5)\qquad L_0\mathbf{u}_1=-2N_1(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)$$

and

$$\begin{aligned}(7.6)\qquad L_0\mathbf{u}_2+H_1L_1\boldsymbol{\phi}_0+N_1(\boldsymbol{\phi}_0,u_1)+N_1(\boldsymbol{\phi}_0,\boldsymbol{\phi}_1)+N_1(\boldsymbol{\phi}_1,\boldsymbol{\phi}_0)\\+N_1(\mathbf{u}_1,\boldsymbol{\phi}_0)+3N_2(\boldsymbol{\phi}_0,\boldsymbol{\phi}_0,\boldsymbol{\phi}_0)-\sigma_2\boldsymbol{\phi}_0=\mathbf{0}.\end{aligned}$$

Combining (7.2) and (7.5) yields

$$\mathbf{u}_1=2\boldsymbol{\phi}_1.$$

This allows us to rewrite (7.6) as

$$(7.7)\qquad L_0\mathbf{u}_2+H_1L_1\phi_0+3(N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)+N_3(\phi_0,\phi_0,\phi_0))=\sigma_2\phi_0.$$

Using the expression for H_1 and operating by $[\cdot]$ on (7.7) produces the following expression for σ_2:

$$\sigma_2=-2H_1[L_1\phi_0].$$

This expression gives the same stability conditions on nonzero H_1 as obtained for H_0.

The above analysis assumes the expansions for σ and $\mathbf{u}$ are convergent. We now prove the local analyticity of these expansions.

Let $\mathbf{u}=\phi_0+\xi$ where $[\xi]=0$. Then, we use the Lyapunov–Schmidt equation associated with $F_\phi(H,\phi)\mathbf{u}=0$ to define F:

$$T\quad:\mathbb{R}\times(\mathbb{R}\times\hat{B})\to\mathbb{R}\times\hat{B}$$

$$:\begin{pmatrix}\varepsilon\\ \sigma\\ \xi\end{pmatrix}\to\begin{pmatrix}(H-H_c)[L_1(\phi_0+\xi)]-\sigma+[\text{terms involving }\varepsilon]\\ \xi+K\{(H-H_c)L_c(\phi_0+\xi)-\sigma(\phi_0+\xi)+\text{terms involving }\varepsilon\}\end{pmatrix}$$

where H is a function of ε such that $H=H_c$ when $\varepsilon=0$. In this case $T(0,0,\mathbf{0})=\mathbf{0}$ and

$$T_{(\sigma,\xi)}(0,0,\mathbf{0})=\begin{pmatrix}-1 & 0\\ -K\phi_0 & I\end{pmatrix}$$

is continuously invertible in $\mathbb{R}\times\hat{B}$, so IFT and the analyticity of F provide convergence of $\xi(\varepsilon)$ and $\sigma(\varepsilon)$ for small $|\varepsilon|$.

This completes the proof of the stability theorem.

8. Results—rolls. The existence and stability theorems provide results for several types of perturbations. Each class of perturbations corresponds to a choice of R and a related function Tr satisfying conditions (4.1):

$$\nabla^2\mathrm{Tr}=-\omega^2\mathrm{Tr}\quad\text{in } R,\qquad\frac{\partial\,\mathrm{Tr}}{\partial n}=0\quad\text{on }\partial R.$$

In general, according to the stability theorem, we must consider $H_0=-[N_1(\phi_0,\phi_0)]/[L_1\phi_0]$ where

$$(8.1)\qquad [N_1(\phi_0,\phi_0)]=\frac{\mu+1}{H_c}(1+e^{-2\omega D})\cdot\left\{\left(1-5e^{-2\omega D}+e^{-4\omega D}+\frac{5}{4}\frac{1-e^{-4\omega D}}{\omega D}\right)\omega^2\int\mathrm{Tr}^3\,d\sigma\right.$$

$$+\frac{1}{2}\left(-3+2e^{-2\omega D}-3e^{-4\omega D}+\frac{1-e^{-4\omega D}}{\omega D}\right.$$

$$\left.\left.+\frac{\mu+1}{\mu-1}2(1+e^{-2\omega D})^2\right)\int|\nabla\,\mathrm{Tr}|^2\mathrm{Tr}\,d\sigma\right\}.$$

In the case of rolls ($R=[-\frac{\pi}{\omega},\frac{\pi}{\omega}]\times[0,1]$ and $\mathrm{Tr}=\cos\omega x_1$) and rectangles ($R=[-\frac{\pi}{\omega},\frac{\pi}{\omega}]\times[-\frac{\pi}{\omega},\frac{\pi}{\omega}]$ and $\mathrm{Tr}=\frac{1}{2}(\cos\omega x_1+\cos\omega x_2)$), $[N_1(\phi_0,\phi_0)]$ is zero. This can be seen by calculating the above expression.

Then, according to Theorem 5 we must next consider H_1 which involves finding ϕ_1. To find ϕ_1 is computationally an extremely difficult problem. One approach is to compute ϕ_1, and hence H_1, numerically. We have decided upon another alternative,

that of calculating ϕ_1 and H_1 for infinite D. Assuming D infinite is logical since all of the experiments and the assumptions we have made are for very large D.

To further justify allowing D to become infinite for the stability calculation we prove, below, Theorem 6 which states that H_1 with $D=\infty$, $H_{1\infty}$, is $O(\frac{1}{D})$ of H_1. Thus, given any combination of magnetic parameters, there is a D so that for depths greater than D the stability of the finite problem is the same as that indicated by the infinite problem.

Let

$$\phi'_{0\infty}=-\mu e^{-\omega x_3}\mathrm{Tr},\quad \phi_{0\infty}=e^{\omega x_3}\mathrm{Tr},\quad z_{0\infty}=\frac{\mu+1}{\mu-1}H_{c\infty}\mathrm{Tr},\quad H^2_{c\infty}=\frac{\mu+1}{\mu-1}\frac{2\omega\tau}{\mu\mu_0}$$

and denote by $\phi_{1\infty}$ and $H_{1\infty}$ the values calculated using (7.2) with ϕ_0 replaced by $\phi_{0\infty}$ in the space $B_{2+\alpha}$ defined in §3 with $D=\infty$.

THEOREM 6.

$$H_1=H_{1\infty}+O\left(\frac{1}{D}\right).$$

Proof. Write ϕ_0 and N_1 as

$$\phi_0=\phi_{00}+e^{-2\omega D}\phi_{01}$$

and

$$N_1(\cdot,\cdot)=N_{10}(\cdot,\cdot)+\frac{1}{D}N_{11}(\cdot,\cdot),$$

where

$$\phi_{00}=\begin{pmatrix}-\mu e^{-\omega x_3}\mathrm{Tr}\\ e^{\omega x_3}\mathrm{Tr}\\ \frac{\mu+1}{\mu-1}\frac{1}{H_c}\mathrm{Tr}\\ \omega\,\mathrm{Tr}\\ (\mu-1)\omega\,\mathrm{Tr}\end{pmatrix},\qquad \phi_{01}=\begin{pmatrix}-\mu e^{\omega x_3}\mathrm{Tr}\\ e^{-\omega x_3}\mathrm{Tr}\\ \frac{\mu+1}{\mu-1}\frac{1}{H_c}\mathrm{Tr}\\ -4\omega\,\mathrm{Tr}\\ -(\mu-1)\omega\,\mathrm{Tr}\end{pmatrix},$$

$$N_{11}(\phi,\phi)=\begin{pmatrix}-2z\frac{\partial^2\phi'}{\partial x_3^2}\\ 2z\frac{\partial^2\phi}{\partial x_3^2}\\ 0\\ 0\\ 0\end{pmatrix},\qquad N_{10}(\cdot,\cdot)=N_1(\cdot,\cdot)-\frac{1}{D}N_{11}(\cdot,\cdot).$$

Note that in the expression for $z_{0\infty}$ contained $H_{c\infty}$ while ϕ_{00} and the expressions not generally depending on D contain an H_c (which does depend on D). This is for convenience and is correct since as D grows large, H_c approaches $H_{c\infty}$, which is constant. Hence, the appearance of H_c in an expression will never affect any order argument concerning the expression.

Recall from Theorem 5 that ϕ_1 satisfies

$$L_0\phi_1=-N_1(\phi_0,\phi_0).$$

For notational covenience, denote $\phi_{1\infty}$ by ϕ_{11} and note that, in this notation, ϕ_{11} must satisfy

$$L_0\phi_{11} = -N_{11}(\phi_{00}, \phi_{00}). \tag{8.2}$$

Then, if ϕ_{1i}, $i=2,3,4,5,6$, are defined to be solutions to the problems

$$L_0\phi_{12} = -e^{-2\omega D}(N_{10}(\phi_{00}, \phi_{01}) + N_{10}(\phi_{01}, \phi_{00})), \tag{8.3}$$

$$L_0\phi_{13} = -e^{-4\omega D}N_{10}(\phi_{01}, \phi_{01}), \tag{8.4}$$

$$L_0\phi_{14} = \frac{1}{D}N_{11}(\phi_{00}, \phi_{00}), \tag{8.5}$$

$$L_0\phi_{15} = \frac{1}{D}e^{-2\omega D}(N_{11}(\phi_{00}, \phi_{01}) + N_{11}(\phi_{01}, \phi_{00})) \quad \text{and} \tag{8.6}$$

$$L_0\phi_{16} = \frac{1}{D}e^{-4\omega D}N_{11}(\phi_{01}, \phi_{01}), \tag{8.7}$$

respectively, we have

$$\phi_1 = \sum_{m=1}^{6} \phi_{1m}.$$

Equations (8.2)–(8.7) are quite similar if we multiply them by -1, $-e^{2\omega D}, e^{4\omega D}$, $D, De^{2\omega D}$, and $De^{4\omega D}$ respectively and absorb this expression into the unknowns ϕ_{11} to ϕ_{16}, respectively. All the above expressions are special cases of the general equation:

$$L_0\phi = N(\phi_{00}, \phi_{01}) \tag{8.8}$$

where

$$N(\phi_{00}, \phi_{01}) = \begin{bmatrix} \sum_{m=0,2,4} \left[K_{10m}\frac{e^{-\omega x_3}}{H_c} + K_{11m}\frac{e^{\omega x_3}}{H_c} + K_{12m}x_3\frac{e^{-\omega x_3}}{H_c} + K_{13m}x_3\frac{e^{\omega x_3}}{H_c} \right] \mathrm{Tr}_m \\ \sum_{m=0,2,4} \left[K_{20m}\frac{e^{\omega x_3}}{H_c} + K_{21m}\frac{e^{-\omega x_3}}{H_c} + K_{22m}x_3\frac{e^{\omega x_3}}{H_c} + K_{23m}x_3\frac{e^{-\omega x_3}}{H_c} \right] \mathrm{Tr}_m \\ \sum_{m=0,2,4} K_{30m}\mathrm{Tr}_m \\ 0 \\ \sum_{m=0,2,4} K_{50m}\mathrm{Tr}_m \end{bmatrix}$$

and, in the case of rolls,

$$\mathrm{Tr}_0 = 1, \quad \mathrm{Tr}_2 = 0, \quad \mathrm{Tr}_4 = \cos 2\omega x_1,$$

or rectangles,

$$\mathrm{Tr} = 1, \quad \mathrm{Tr}_2 = \cos\omega x_1 \cos\omega x_2, \quad \mathrm{Tr}_4 = \tfrac{1}{2}(\cos 2\omega x_1 + \cos 2\omega x_2).$$

Then as D varies, K_{ijm} remains constant. For example, if we consider (8.5), the right-hand side of the equation is

$$N_{11}(\phi_{00},\phi_{00})=\begin{bmatrix} -\frac{\mu+1}{\mu-1}\mu\omega^3\,\mathrm{Tr}^2\,\frac{x_3e^{-\omega x_3}}{H_c}+2\,\frac{\mu+1}{\mu-1}\mu\omega^2\,\mathrm{Tr}^2\,\frac{e^{-\omega x_3}}{H_c} \\ \frac{\mu+1}{\mu-1}\omega^3\,\mathrm{Tr}^2\,\frac{x_3e^{\omega x_3}}{H_c}-2\,\frac{\mu+1}{\mu-1}\omega^2\,\mathrm{Tr}^2\,\frac{e^{\omega x_3}}{H_c} \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

so K_{10m}, K_{12m}, K_{20m}, and K_{22m} for $m=0,2$, and 4 are the only nonzero coefficients in N.

To be able to analyze the order of growth of H_1, it is necessary to know which of the K_{ijm} are identically zero. This can be done by evaluating each of the right-hand sides of equations (8.2)–(8.7). The results of these calculations are presented in Table 1.

TABLE 1

Coefficients K_{ijm} which are zero (0) or nonzero (–) in the operator N of (8.8).

	K_{ijm}									
Equation	$ijm=10m$	$11m$	$12m$	$13m$	$20m$	$21m$	$22m$	$23m$	$30m$	$50m$
(8.1)	–	0	0	0	–	0	0	0	–	–
(8.2)	–	–	0	0	–	–	0	0	–	–
(8.3)	0	–	0	0	0	–	0	0	–	–
(8.4)	–	0	–	0	–	0	–	0	0	0
(8.5)	–	–	–	–	–	–	–	–	0	0
(8.6)	0	–	0	–	0	–	0	–	0	0

Next, given the general expression for N, we can solve (8.8) for a function ϕ of the form

$$\phi=\begin{bmatrix} \phi' \\ \phi \\ z \\ f(\phi',\phi) \\ g(\phi',\phi) \end{bmatrix}$$

where f and g are determined by the boundary conditions,

$$\phi'=\sum_{m=0,2,4}\left(P_{10m}e^{-\omega x_3}+P_{11m}e^{\omega x_3}+P_{12m}x_3e^{-\omega x_3}+P_{13m}x_3e^{\omega x_3}\right)\mathrm{Tr}_m$$
$$+\sum_{m=2,4}\left(P_{14m}e^{-\omega\sqrt{m}x_3}+P_{15m}e^{\omega\sqrt{m}x_3}\right)\mathrm{Tr}_m+x_3P_{140}+P_{150},$$
$$\phi=\sum_{m=0,2,4}\left(P_{20m}e^{\omega x_3}+P_{21m}e^{-\omega x_3}+P_{22m}x_3e^{\omega x_3}+P_{23m}x_3e^{-\omega x_3}\right)\mathrm{Tr}_m$$
$$+\sum_{m=2,4}\left(P_{24m}e^{\omega\sqrt{m}x_3}+P_{25m}e^{-\omega\sqrt{m}x_3}\right)\mathrm{Tr}_m+x_3P_{240}+P_{250},$$

and

$$z=\frac{1}{(\mu-1)H_c}(\phi'-\phi).$$

When L_0 is applied to ϕ and the result is set equal to N, certain coefficients, P_{ijm}, can be found explicitly. It is not hard to see that

$$P_{i0m}=\frac{(K_{10m}/H_c+(2i-1)2K_{12m}/\omega(m-1)H_c)}{\omega^2(m-1)},$$

$$P_{i1m}=\frac{(K_{11m}/H_c-(2i-1)2K_{13m}/\omega(m-1)H_c)}{\omega^2(m-1)},$$

$$P_{i2m}=\frac{K_{12m}}{\omega^2(m-1)H_c} \quad \text{and}$$

$$P_{i3m}=\frac{K_{13m}}{\omega^2(m-1)H_c}$$

for $i=1,2$ and $m=0,2,4$. The coefficients which are identically zero are given in Table 2.

The remaining coefficients are defined implicitly in terms of D,ω and K_{ijm}, $i=1,2$, $j=0,1,2,3$, and $m=0,2,4$, by the four equations that were not readily solvable. These equations are linear in the unknowns P_{ijm} and, since the K_{ijm}'s are constant, the order of growth of the numerator and denominator for each coefficient can be determined by Cramer's rule. The results of this process are given in Table 3.

For example, if we again consider equation (8.5), then $-P_{i44}$ and P_{i54}, $i=1,2$, are estimated from the following equation:

$$\begin{bmatrix} -2\omega e^{-\omega 2D} & 0 & 2\omega e^{\omega 2D} & 0 \\ 0 & 2\omega e^{-\omega 2D} & 0 & -2\omega e^{\omega 2D} \\ \omega & 2\omega & \mu\omega & 2\mu\omega \\ K_1-K_2 & K_1-2K_2 & -K_1-K_2 & -K_1-2K_2 \end{bmatrix} \begin{bmatrix} P_{144} \\ P_{244} \\ P_{154} \\ P_{254} \end{bmatrix}$$

$$= \begin{bmatrix} \omega e^{-\omega D}P_{104}+\omega DP_{124}e^{-\omega D} \\ -\omega e^{-\omega D}P_{204}+\omega DP_{244}e^{-\omega D} \\ \omega P_{104}-P_{124}+\mu P_{204}-\mu\omega P_{224} \\ -K_1(P_{104}-P_{204})-K_2\left(P_{104}+P_{204}-\dfrac{P_{124}-P_{224}}{\omega}\right) \end{bmatrix}$$

where

$$K_1=\frac{\dfrac{\Delta\rho g}{\tau}+4\omega^2}{(\mu-1)H_c}$$

and

$$K_2=\frac{\mu\mu 0}{\tau}H_c\omega.$$

If we write the above equation as

$$M\mathbf{P}=\mathbf{R}$$

TABLE 2

Coefficients P_{ijm} which are determined explicitly to be zero (0) or nonzero (−).

Equation	P_{ijm}					
	$ijm=i0m$	$i1m$	$i2m$	$i3m$	$i4m$	$i5m$
(8.1)	−	0	0	0	−	−
(8.2)	−	−	0	0	−	−
(8.3)	0	−	0	0	−	−
(8.4)	−	0	−	0	−	−
(8.5)	−	−	−	−	−	−
(8.6)	0	−	0	−	−	−

TABLE 3

Order of growth of numerator and denominator of the implicitly determined P_{ijm} coefficients.

Equation	Numerator/denominator, P_{ijm}					
	$ijm=i40$	$i50$	$i42$	$i52$	$i44$	$i54$
(8.1)	$e^{-\omega D}/1$	$e^{-\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$1/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$1/e^{4\omega D}$
(8.2)	$e^{\omega D}/1$	$e^{\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$e^{(1+\sqrt{2})\omega D}/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$e^{2\omega D}/e^{4\omega D}$
(8.3)	$e^{\omega D}/1$	$e^{\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$e^{(1+\sqrt{2})\omega D}/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$e^{2\omega D}/e^{4\omega D}$
(8.4)	$De^{-\omega D}/1$	$De^{-\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$1/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$1/e^{4\omega D}$
(8.5)	$De^{\omega D}/1$	$De^{\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$e^{(1+\sqrt{2})\omega D}/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$e^{2\omega D}/e^{4\omega D}$
(8.6)	$De^{\omega D}/1$	$De^{\omega D}/1$	$e^{2\sqrt{2}\omega D}/e^{2\sqrt{2}\omega D}$	$e^{(1+\sqrt{2})\omega D}/e^{2\sqrt{2}\omega D}$	$e^{4\omega D}/e^{4\omega D}$	$e^{2\omega D}/e^{4\omega D}$

and denote the matrix defined by replacing the first column of M by the vector $\mathbf{R}$ by N, then

$$\det M = 4\omega^3 K_1 e^{4\omega D}\left[1+\mu e^{-4\omega D}(1-e^{-4\omega D})-e^{-4\omega D}\right]$$
$$\geq 2\omega^3 K_1 e^{4\omega D} \quad \text{(for sufficiently large } D\text{)},$$
$$\det N = O(e^{4\omega D}),$$

and

$$P_{144} = O(1).$$

It is not difficult to use the growth conditions on the coefficients P_{ijm} to give growth conditions on the functions $\phi', \phi, \partial\phi'/\partial x_3, \partial\phi/\partial x_3, \partial^2\phi'/\partial x_3^2$ and $\partial^2\phi/\partial x_3^2$. These results are given in Table 4.

TABLE 4

The functions shown represent the lowest order terms in expressions for ϕ'_{1m}, ϕ_{1m}, their first and second derivatives with respect to x_3.

	$m=1$	2	3	4	5	6
ϕ'	$e^{-\omega x_3}$	$e^{-\omega D}$	$e^{-3\omega D}$	$\frac{x_3}{D}e^{-\omega x_3}$	$x_3 e^{-\omega D}$	$e^{-3\omega D}$
$\partial\phi'/\partial x_3$	$e^{-\omega x_3}$	$e^{-\omega D}$	$e^{\omega(x_3-2D)}$	$\frac{x_3}{D}e^{-\omega x_3}$	$e^{-\omega D}$	$\frac{x_3}{D}e^{\omega(x_3-4D)}$
$\partial^2\phi'/x_3^2$	$e^{-\omega x_3}$	$e^{\omega(x_3-2D)}$	$e^{\omega(x_3-2D)}$	$\frac{x_3}{D}e^{-\omega x_3}$	$x_3 e^{\omega(x_3-2D)}$	$\frac{x_3}{D}e^{\omega(x_3-4D)}$
ϕ	$e^{\omega x_3}$	$e^{-\omega D}$	$e^{-3\omega D}$	$\frac{x_3}{D}e^{\omega x_3}$	$x_3 e^{-\omega D}$	$e^{-3\omega D}$
$\partial\phi/\partial x_3$	$e^{\omega x_3}$	$e^{-\omega D}$	$e^{\omega(-x_3-2D)}$	$\frac{x_3}{D}e^{\omega x_3}$	$e^{-\omega D}$	$\frac{x_3}{D}e^{\omega(-x_3-4D)}$
$\partial^2\phi'/\partial x_3^2$	$e^{\omega x_3}$	$e^{\omega(-x_3-2D)}$	$e^{\omega(-x_3-2D)}$	$\frac{x_3}{D}e^{\omega x_3}$	$x_3 e^{\omega(-x_3-2D)}$	$\frac{x_3}{D}e^{\omega(-x_3-4D)}$

And finally, one last step is to estimate H_1. Recall H_1 is given by

$$H_1 = \frac{[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)+N_3(\phi_0,\phi_0,\phi_0)]}{[L_1\phi_0]}.$$

Write the functional $[\cdot]$ as

$$[\cdot] = [\cdot]_1 + e^{-2\omega D}[\cdot]_2,$$

where

$$[\cdot]_1 = \langle \cdot, \mathbf{u}_1^* \rangle, \qquad [\cdot]_2 = \langle \cdot, \mathbf{u}_2^* \rangle,$$

$$\mathbf{u}_1^* = \frac{1}{K}\begin{bmatrix} -e^{-\omega x_3}\mathrm{Tr} \\ e^{\omega x_3}\mathrm{Tr} \\ -\dfrac{\mu+1}{\mu-1}\dfrac{\tau}{\mu\mu_0}\dfrac{1}{H_c}\mathrm{Tr} \\ \omega\,\mathrm{Tr} \\ -\dfrac{2}{\mu-1}\mathrm{Tr} \end{bmatrix}, \qquad \mathbf{u}_2^* = \frac{1}{K}\begin{bmatrix} -e^{\omega x_3}\mathrm{Tr} \\ e^{-\omega x_3}\mathrm{Tr} \\ -\dfrac{\mu+1}{\mu-1}\dfrac{\tau}{\mu\mu_0}\dfrac{1}{H_c}\mathrm{Tr} \\ \omega\,\mathrm{Tr} \\ -\dfrac{2}{\mu-1}\mathrm{Tr} \end{bmatrix}$$

and K is defined in §4. Also, let $[\cdot]_\infty$ denote the limit of $[\cdot]_1$ for infinite D. This limit involves H_c approaching $H_{c\infty}$, $\mathbf{u}^*$ approaching $\mathbf{u}_1^*$ and the space in the x_3 direction becoming infinite. To obtain an estimate of H_1 we write H_1 as

$$H_1 = -\frac{[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)]_1 + [N_3(\phi_0,\phi_0,\phi_0)]_1}{[L_1\phi_0]_1 + e^{-2\omega D}[L_1\phi_0]_2}$$
$$-\frac{e^{-2D}[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)]_2 + [N_3(\phi_0,\phi_0,\phi_0)]_2}{[L_1\phi_0]_1 + e^{-2\omega D}[L_1\phi_0]_2}$$

and proceed to estimate each of the above terms. For example,

$$[L_1\phi_0]_1 = \left\langle \begin{bmatrix} 0 \\ 0 \\ \dfrac{\mu\mu 0}{\tau}\omega(\mu-1)(1-e^{-2\omega D})\mathrm{Tr} \\ \dfrac{(\mu+1)}{H_c}(1+e^{-2\omega D})\mathrm{Tr} \\ 0 \end{bmatrix}, \mathbf{u}_1^* \right\rangle$$

$$= \left\langle \begin{bmatrix} 0 \\ 0 \\ \dfrac{\mu\mu 0}{\tau}\omega(\mu-1)\mathrm{Tr} \\ \dfrac{\mu+1}{H_c}\mathrm{Tr} \\ 0 \end{bmatrix}, \begin{bmatrix} - \\ - \\ \dfrac{\mu+1}{\mu-1}\dfrac{\tau}{\mu\mu 0}\dfrac{\mathrm{Tr}}{H_c} \\ \omega\,\mathrm{Tr} \\ - \end{bmatrix} \right\rangle$$

$$+2\frac{\mu+1}{H_c}\omega(e^{-4\omega D}+e^{-2\omega D})\int \mathrm{Tr}^2$$

$$= [L_1\phi_{00}]_1 + O(e^{-2\omega D}).$$

In this case, since letting $D \to \infty$ in $[L_1\phi_{00}]_1$ involves $H_c \to H_{c\infty}$, it is easy to see that

$$[L_1\phi_0]_1 = [L_1\phi_{00}]_\infty + O(3^{-2\omega D}).$$

Similar estimates give

$$[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)]_1 = [N_1(\phi_{0\infty},\phi_{1\infty})+N_1(\phi_{1\infty},\phi_{0\infty})] + O(De^{-2\omega D}),$$
$$[N_1(\phi_0,\phi_1)+N_1(\phi_1,\phi_0)]_2 = O(D),$$

$$[N_3(\phi_0,\phi_0,\phi_0)]_1 = [N_3(\phi_{0\infty},\phi_{0\infty},\phi_{0\infty})]_\infty + O\left(\frac{1}{D}\right),$$
$$[N_3(\phi_0,\phi_0,\phi_0)]_2 = O(D),$$

$$[L_1\phi_0]_1 = [L_1\phi_{0\infty}]_\infty + O(e^{-2\omega D}),$$
$$[L_1\phi_0]_2 = O(1).$$

It should be noted that several of the above expressions involve integrating over an infinite region instead of the finite region. Each time that this substitution must be made, there are exponentials being integrated which preserve the desired order. Therefore

$$H_1 = \frac{[N_1(\phi_{0\infty},\phi_{0\infty})]_\infty + O(1/D)}{[L_1\phi_{0\infty}]_\infty + O(e^{-2\omega D})} = H_{1\infty} + O(1/D),$$

which is what we were to prove.

Let us now return to the calculation of the approximate stability criterion for solutions in the form of rolls. Using the approximation made in Theorem 5, use the following expressions to determine $\phi_{1\infty}$. When we solve

$$L_0\phi_{1\infty} = -N_1(\phi_{0\infty},\phi_{0\infty}) = \begin{bmatrix} -\mu\dfrac{\mu+1}{\mu-1}\dfrac{1}{H_{c\infty}}\omega^3\left(\dfrac{1}{2}-\dfrac{1}{2}\cos 2\omega x_1\right)e^{-\omega x_3} \\ \dfrac{\mu+1}{\mu-1}\dfrac{1}{H_{c\infty}}\omega^3\left(\dfrac{1}{2}-\dfrac{1}{2}\cos 2\omega x_1\right)e^{\omega x_3} \\ \dfrac{\mu\mu_0}{\tau}\dfrac{\omega^2}{\mu-1}\dfrac{1}{4}\left(2(\mu+1)^2-8\mu\cos 2\omega x_1\right) \\ -\mu\dfrac{\mu+1}{\mu-1}\dfrac{1}{H_{c\infty}}\omega^2(1-\cos 2\omega x_1) \\ 0 \end{bmatrix},$$

we get

$$\phi'_{1\infty} = \mu\frac{\mu\mu_0}{\tau}\frac{H_{c\infty}}{4}(\mu-1)\left\{e^{-\omega x_3}-1+\left(e^{-\omega x_3}-\frac{3\mu-1}{\mu+1}e^{-2\omega x_3}\right)\cos 2\omega x_1\right\},$$

$$\phi_{1\infty} = \frac{\mu\mu_0}{\tau}\frac{H_{c\infty}}{4}(\mu-1)\left\{e^{\omega x_3}-1+\left(e^{\omega x_3}+\frac{\mu-3}{\mu+1}e^{2\omega x_3}\right)\cos 2\omega x_1\right\},$$

$$z_{1\infty} = \frac{\mu\mu_0}{\tau}\frac{(\mu-1)}{2}\cos 2\omega x_1.$$

We then calculate H_1 by the expression

$$(8.9)\quad H_{1\infty} = -\left[N_1(\phi_{0\infty},\phi_{0\infty}) + N_1(\phi_{1\infty},\phi_{0\infty}) + N_3(\phi_{0\infty},\phi_{0\infty},\phi_{0\infty})\right]_\infty / \left[L_1\phi_{0_\infty}\right]_\infty$$
$$= H_{c\infty}\frac{\mu\mu_0}{\tau}\frac{\omega}{8}\frac{1}{\mu+1}\frac{11}{4}(\mu-1)^2+\frac{25}{3}\mu$$

where $H_{c\infty}^2 = 2\omega(\mu+1/(\mu-1)^2)(\tau/\mu\mu_0)$. Since $H_{1\infty}$ is greater than zero for all μ, the bifurcating branch is stable.

To this level of approximation we obtain

$$H \approx H_c + \varepsilon^2 H_1 \quad \text{and} \quad z \approx z_0 + \varepsilon^2 z_1.$$

By combining these two approximations to eliminate ε we find

$$z = \sqrt{\frac{H-H_c}{H_c}}\frac{2(\mu+1)}{\omega\sqrt{h}}\cos\omega x_1 + \frac{H-H_c}{H_c}\frac{2(\mu^2-1)}{\omega h}\cos 2\omega x_1,$$

where $h = \frac{11}{4}(\mu-1)^2+\frac{25}{3}\mu$.

This two-dimensional case has been considered in a previous article by Zaitsev and Shliomis [6], who obtained different results. The brevity of their article makes a comparison difficult.

9. Results—rectangles. In this section a class of perturbations is considered which corresponds to a choice of $R=[-\frac{\pi}{\omega},\frac{\pi}{\omega}]\times[-\frac{\pi}{\omega},\frac{\pi}{\omega}]$ and $\mathrm{Tr}=\frac{1}{2}(\cos\omega x_1+\cos\omega x_2)$. This is the physically observable square relief pattern found quite often in the electrical analogue of the ferrofluid problem. Again as in the case of rolls, $H_0=0$. Also, in this case, H_1 is extremely difficult to calculate. The proof that H_1 is equal to $H_{1\infty}+O(1/D)$ is the same, using Tr, as given above. Thus we again use the approximation that we did in the case of rolls and calculate $H_{1\infty}$ instead of H_1.

With D infinite we find that

$$\phi'_{0\infty} = -\mu e^{-\omega x_3}\mathrm{Tr}, \quad \phi_{0\infty} = e^{\omega x_3}\mathrm{Tr}, \quad z_{0\infty} = \frac{\mu+1}{\mu-1}\frac{1}{H_{c\infty}}\mathrm{Tr}.$$

When $L_0\phi_{1\infty} = -\mathrm{N}_{11}(\phi_{0\infty},\phi_{0\infty})$ is solved, we find that

$$\phi'_{1\infty} = \mu\frac{\mu\mu_0}{\tau}\frac{(\mu-1)}{4}H_c\left\{\frac{1}{2}(e^{-\omega x_3}-1)+\left(e^{-\omega x_3}-\frac{6(\mu-1)+(4\mu-3)\sqrt{2}}{\mu+1}e^{-\sqrt{2}\omega x_3}\right)\mathrm{Tr}_2\right.$$
$$\left.+\frac{1}{2}\left(e^{-\omega x_3}-\frac{3\mu-1}{\mu+1}e^{-2\omega x_3}\right)\mathrm{Tr}_4\right\},$$

$$\phi_{1\infty} = \frac{\mu\mu_0}{\tau}\frac{(\mu-1)}{4}H_c\left\{\frac{1}{2}(e^{\omega x_3}-1)+\left(e^{\omega x_3}+\frac{6(\mu-1)+(3\mu-4)\sqrt{2}}{\mu+1}e^{\sqrt{2}\omega x_3}\right)\mathrm{Tr}_2\right.$$
$$\left.+\frac{1}{2}\left(e^{\omega x_3}+\frac{\mu-3}{\mu+1}e^{2\omega x_3}\right)\mathrm{Tr}_4\right\},$$

$$z_{1\infty} = \frac{\mu\mu_0}{\tau}\frac{(\mu-1)}{4}\left\{\left(5+4\sqrt{2}\right)\mathrm{Tr}_2+\mathrm{Tr}_4\right\},$$

where $\mathrm{Tr}_2=\cos\omega x_1\cos\omega x_2$ and $\mathrm{Tr}_4=\frac{1}{2}(\cos 2\omega x_1+\cos 2\omega x_2)$. Then the appropriate stability criterion H_1 can be calculated using the form of (8.9), and we get

$$H_{1\infty}=-H_{c\infty}\frac{\mu\mu_0}{\tau}\frac{\omega}{6}\frac{1}{\mu+1}h$$

where $h=(\frac{361}{64}+2\sqrt{2})(\mu-1)^2-(\frac{7}{16}+\frac{\sqrt{2}}{2})\mu$.

In this case H_1 is positive if μ remains between approximately $.69^+$ and 1.44^+ where the branch is stable. Outside this range the branch is locally unstable. To this degree of approximation the expression for z in the stable region is

$$z=\sqrt{\frac{H-H_c}{H_c}}\frac{(\mu+1)\sqrt{3}}{2\omega}(-h)^{-1/2}\mathrm{Tr}$$

$$-\frac{H-H_c}{H_c}\frac{3}{2}\frac{\mu^2-1}{\omega}h^{-1}\left(\left(5+4\sqrt{2}\right)\mathrm{Tr}_2+\mathrm{Tr}_4\right).$$

We note that the stability results for finite D would follow in an analogous manner. However, the amount of work necessary is prohibitive.

No previous results for this case had been obtained.

10. Hexagons. In this section we consider the final class of perturbation relief patterns: hexagons. Set $R=\{(x_1,x_2)\,|\,|x_2|\leq 2\pi/\omega\sqrt{3},\ |x_1-\sqrt{3}\,x_2|\leq 2\pi/\omega$ and $|x_1+\sqrt{3}\,x_2|\leq 2\pi/\omega\}$ with $\mathrm{Tr}=\frac{1}{3}(\cos\omega(x_1+\sqrt{3}\,x_2))/2+\cos\omega(x_1-\sqrt{3}\,x_2)/2+\cos\omega x_1$. As we shall see below, H_0 is not zero in this case. Hence we have an analytic result and need not resort to an order argument. We obtain

$$\phi_0'=-\mu\,\mathrm{Tr}(e^{-\omega x_3}+e^{-\omega(2D-x_3)}),$$

$$\phi_0=\mathrm{Tr}(e^{\omega x_3}+e^{-\omega(2D+x_3)}),$$

$$z_0=\frac{\mu+1}{\mu-1}\frac{1}{H_c}\mathrm{Tr}(1+e^{-2\omega D}).$$

If we define

$$\mathrm{Tr}_0=1,$$

$$\mathrm{Tr}_1=\mathrm{Tr},$$

$$\mathrm{Tr}_3=\cos\frac{3\omega x_1}{2}\cos\frac{\omega\sqrt{3}\,x_2}{2}+\frac{1}{2}\cos\omega\sqrt{3}\,x_2,$$

$$\mathrm{Tr}_4=\cos\omega x_1\cos\omega\sqrt{3}\,x_2+\frac{1}{2}\cos 2\omega x_1,$$

then

$$\mathrm{Tr}^2=\mathrm{Tr}_1^2=\frac{1}{9}\left(\frac{3}{2}\mathrm{Tr}_0+2\,\mathrm{Tr}_1+2\,\mathrm{Tr}_3+\mathrm{Tr}_4\right),$$

$$|\nabla\,\mathrm{Tr}|^2=\frac{\omega^2}{9}\left(\frac{3}{2}\mathrm{Tr}_0+\mathrm{Tr}_1-\mathrm{Tr}_3-\mathrm{Tr}_4\right),$$

$$\int_R\mathrm{Tr}_i\mathrm{Tr}_j=0\quad\text{for } i\neq j.$$

Using these facts, in this case (8.1) can be simplified to

$$-[N(\phi_0,\phi_0)] = -\frac{\mu+1}{18H_c}\frac{\omega^2}{k}(1+e^{-2\omega D}) \cdot \left(1-22e^{-2\omega D}+e^{-4\omega D} -8/\omega D(1-e^{-4\omega D})+\frac{\mu+1}{\mu-1}(1+e^{-2\omega D})^2\right)\int_R \mathrm{Tr}^2\, d\sigma.$$

Also

$$[L_1\phi_0] = 2\frac{\mu+1}{H_c}\frac{\omega}{K}(1-e^{-2\omega D})\int_R \mathrm{Tr}^2\, d\sigma.$$

Hence

$$H_0 = -\frac{\omega}{36}\coth\omega D\left[1-22e^{-2\omega D}+e^{-4\omega D}-8/\omega D(1-e^{-4\omega D})+\frac{\mu+1}{\mu-1}(1+e^{-2\omega D})^2\right].$$

Since H_0 is less than zero for large D, the branch is unstable. This agrees with the result obtained by Kuznetsov and Spector [2] and Gailitis [5] who considered the case of infinite D.

11. Summary. Under the assumptions applied in this article, we have proved in §6 that the operator F defined in §3 has, in addition to the trivial branch of solutions $(\mathbf{0}, H)$, another branch emanating from $(\mathbf{0}, H_c)$. This branch is locally analytic, and stability conditions along the branch are obtained in §7. These results agree with assumptions made by previous authors who used fluids of infinite depth.

When the stability conditions were applied to three selected periodic forms, qualitative results were obtained. The simplest of the periodic forms to analyze, hexagons, produced no local stable branch of solutions. The results have also been obtained elsewhere with fluids of infinite depth.

The other two analyses required simplification of the equations. The resulting stability criterion is equal to the desired stability criterion plus a term which is $O(\frac{1}{D})$. The branch of the square periodic pattern of nontrivial solutions was found to be stable or unstable locally as a function of the magnetic permeability. No previous results had been obtained for this case. As for the rolls structure, the local branch was found always to be stable in a neighborhood of the trivial solution. This disagrees with the results previously reported in an infinitely deep fluid.

The major assumptions we have made for this analysis which apply to the physical situation are the following. The trivial solution must be hydrodynamically stable; the magnetic permeability of the ferrofluid must remain constant throughout the range of the magnetic field strength under consideration; the fluid carries no electric charge; the fluid has a large depth; the magnetic field is initially directed vertically, and all solutions are static.

REFERENCES

[1] M. D. Cowley and R. E. Rosensweig, *The interfacial stability of a ferromagnetic fluid*, J. Fluid Mech., 30 part 4 (1967), pp. 671–688.

[2] E. A. Kuznetsov and M. D. Spector, *Existence of a hexagonal relief on the surface of a dielectric fluid in an external electric fluid*, Zh. Eksper. Teoret. Fiz., 71 (1976), pp. 262–272. (In Russian.)

[3] R. MOSKOWITZ, *Ferrofluids: Liquid magnetics*, IEEE Spectrum, March, 1975.
[4] D. H. SATTINGER, *Topics in stability and bifurcation theory*, Lecture Notes in Mathematics 309, Springer-Verlag, Berlin, 1973.
[5] A. GAILITIS, *Form of surface stability of a ferromagnetic fluid*, Magnit. Gidrodinamika, 5 (1969), pp. 68–70. (In Russian.)
[6] V. M. ZAITSEV AND M. I. SHLIOMIS, *Nature of the instability of the interface between two liquids in a constant field*, Dokl. Akad. Nauk SSSR 188 (1969), p. 1261.
[7] YU. D. BARKOV AND V. G. BASHTOVOI, *Experimental investigation of instability in a plane layer of magnetizable fluid*, Magnit. Gidrodinamika, 3, 4 (1977), pp. 137–140.
[8] A. GAILITIS, *Formation of hexagonal pattern on surface of a ferromagnetic fluid in an applied magnetic field*, J. Fluid Mech., 82 (1977), pp. 401–413.
[9] V. G. BASHTOVOI, *Instability of thin layer of magnetizable liquid with two free boundaries*, Magnit. Gidrodinamika, 3, 3 (1977), pp. 23–29.
[10] R. E. ZELAZO AND J. R. MELCHER, *Dynamics and stability of ferrofluids: Surface interactions*, J. Fluid. Mech., 39 part 1 (1969), pp. 1–24.
[11] M. ZAHM AND K. E. SHENTON, *Magnetic fluids bibliography*, IEEE Trans. Magnetics, 16 (1980), pp. 387–415.
[12] D. D. JOSEPH AND R. I. FOSDICK, *The free surface on a liquid between cylinders rotating at different speeds*, Part I, Arch. Rational Mech. Anal., 49 (1973), pp. 321–380.
[13] D. H. SATTINGER, *On the free surface of a viscous fluid motion*, preprint.
[14] _____, *Group representation theory, bifurcation theory and pattern formation*, J. Functional Anal. 28 (1978), pp. 58–101.
[15] E. H. BOGARDUS, R. SCRANTON AND D. A. THOMAS, *Pulse magnetization measurements in ferrofluids*, IEEE Trans. Magnetics, 11 (1975), pp. 1364–1366.
[16] E. A. PETERSEN, *Concentration Effects in Ferrofluids*, Doctoral Thesis, Colorado State University, Fort Collins, CO, 1975.
[17] D. D. JEFIMENKO, *Electricity and Magnetism*, Appleton-Century-Crofts, New York, 1966.
[18] J. V. WEHAUSEN AND E. V. LAITARE, *Surface waves*, Handbuch Der Physik, Vol. IX, Fluid Dynamics III, S. Flügge, ed., Springer-Verlag, Berlin, 1960.
[19] I. S. SOKOLNIKOFF, *Tensor Analysis, Theory and Applications to Geometry and Mechanics of Continua*, 2nd Ed., John Wiley, New York, 1964.
[20] C. W. MISNER, K. S. THORNE AND J. A. WHEELER, *Gravitation*, W. H. Freeman, San Francisco, 1973.
[21] D. GILBARG AND N. S. TRUDINGER, *Elliptical Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
[22] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0009 $01.25/0

A CHARACTERIZATION OF THE RANGE OF THE DIVERGENT BEAM x-RAY TRANSFORM*

DAVID V. FINCH† AND DONALD C. SOLMON†

Abstract. In this paper we give a characterization of the range of the divergent beam x-ray transform when the source set is a sphere. The result is analogous to the theorem of Helgason [Acta. Math., 113 (1965), pp. 153–180] and Ludwig [Comm. Pure Appl. Math., 69 (1966), pp. 49–81] on the range of the Radon transform.

1. Introduction. For $a\in R^n$ and $\theta\in S^{n-1}$, the divergent beam x-ray transform of a function f from the source point a in the direction θ is defined by

$$\mathcal{D}_a f(\theta)=\int_0^\infty f(a+t\theta)\,dt. \tag{1.1}$$

Physically, one thinks of a as the x-ray source and θ as the direction of the photon beam.

The main result of the paper is a characterization of the range of $\mathcal{D}$ when the set of sources is a sphere. The characterization is similar to and is derived from the Helgason–Ludwig theorem on the range of the Radon transform.

2. The parallel beam x-ray transform. It is assumed throughout that Ω is a bounded, open, convex subset of n-dimensional Euclidean space R^n, $n\geq 2$, with closure $\overline{\Omega}$. Unless otherwise stated f is a square integrable function vanishing outside of Ω, i.e., $f\in L_0^2(\Omega)$.

The parallel beam x-ray transform of f in the direction θ at the point x in $\theta^\perp$ (the subspace orthogonal to θ) is defined by

$$\mathcal{P}_\theta f(x)=\mathcal{P}f(\theta,x)=\int_{-\infty}^\infty f(x+t\theta)\,dt. \tag{2.1}$$

(Here one can think of θ as the direction of the photon beam and x as a point on the x-ray film.) The parallel beam x-ray transform of a function f is a function $\mathcal{P}f$ on $T(S^{n-1})=\{(\theta,x):\theta\in S^{n-1}$ and $\langle x,\theta\rangle=0\}$. This may be identified with the tangent bundle to the sphere. When $n=2$, apart from notation, the parallel beam x-ray transform is the Radon transform.

Remark 2.2. In the initial device for computed tomography [4] a parallel x-ray beam was used and two-dimensional cross-sections were reconstructed. In the second generation of scanners two-dimensional cross-sections are still reconstructed, but a two-dimensional divergent beam is used in place of the parallel beam in order to allow faster scan times.

With the Fourier transform on R^m defined by

$$\hat{f}(\xi)=(2\pi)^{-m/2}\int_{R^m} f(x)e^{-i\langle x,\xi\rangle}\,dx$$

one readily obtains the so called "central section theorem,"

$$\begin{aligned}(\mathcal{P}_\theta f)\hat{}\,(\xi)&=(2\pi)^{(1-n)/2}\int_{\theta^\perp}\mathcal{P}_\theta f(x)e^{-i\langle x,\xi\rangle}\,dx\\&=(2\pi)^{1/2}\hat{f}(\xi),\quad\text{where }\langle\theta,\xi\rangle=0.\end{aligned} \tag{2.3}$$

*Received by the editors October 6, 1981, and in revised form March 4, 1982. This research was supported by the National Science Foundation under grant MCS 8101586.

†Department of Mathematics, Oregon State University, Corvallis, Oregon 97331.

The Sobolev space $H^s(R^m)$, $s \geq 0$, consists of the measurable functions u on R^m such that

$$\|u\|^2_{H^s(R^m)} = \int_{R^m} |\hat{u}(\xi)|^2 \left(1+|\xi|^2\right)^s d\xi < \infty.$$

The Sobolev space $H^{0,s}(T(S^{n-1}))$ is obtained by putting the Euclidean Sobolev space s norm on each fiber, so that

$$\|g\|^2_{H^{0,s}(T(S^{n-1}))} = \int_{S^{n-1}} \|g(\theta,\cdot)\|^2_{H^s(\theta^\perp)} d\theta.$$

The following result, which for $n=2$ is contained in the theorems of Helgason [1] and Ludwig [7] on the range of the Radon transform and for $n>2$ is contained in the results of [11], will be needed. See also Helgason [2].

THEOREM 2.4. *The map $\mathcal{P}: L^2_0(\Omega) \to H^{0,1/2}(T(S^{n-1}))$ is continuous and has a continuous inverse. Moreover, if g is a measurable function on $T(S^{n-1})$ then $g = \mathcal{P}f$ for some $f \in L^2_0(\Omega)$ if and only if*

(i) $g \in H^{0,1/2}(T(S^{n-1}))$,

(ii) $g(\theta,x)=0$ *if the line through x in the direction θ misses Ω,*

(iii) *for each nonnegative integer m and each $\xi \in R^n$ and θ with $\langle \theta,\xi\rangle = 0$, $p_{m,\theta}(\xi) = \int_{\theta^\perp} \langle x,\xi\rangle^m g(\theta,x)\,dx$ is independent of the choice of θ and the function $p_m(\xi) = p_{m,\theta}(\xi)$, which is therefore well defined on R^n, is a homogeneous polynomial of degree $\leq m$.*

Remark 2.5. The preceding result remains valid if $L^2_0(\Omega)$ is replaced by $H^s_0(\Omega)$, $s \geq 0$, provided that $H^{0,1/2}(T(S^{n-1}))$ is replaced by $H^{0,s+1/2}(T(S^{n-1}))$. Here $H^s_0(\Omega)$ consists of those functions in $H^s(R^n)$ which vanish outside of Ω. This extension follows immediately from Theorem 2.4 and the identity [10, p. 1256]

$$\|\Lambda^{s+1/2}\mathcal{P}f\|^2_{L^2(T(S^{n-1}))} = 2\pi|S^{n-2}|\,\|\Lambda^s f\|^2_{L^2(R^n)},$$

Λ being the operator of Calderon defined in terms of the Fourier transform by

$$(\Lambda f)\hat{}\,(\xi) = |\xi|\hat{f}(\xi) \tag{2.6}$$

and

$$|S^{n-2}| = \frac{2\pi^{(n-1)/2}}{\Gamma((n-1)/2)}, \tag{2.7}$$

the surface area of the $(n-2)$ sphere.

The Helgason–Ludwig theorem ($n=2$) plays a central role in the procedure of Louis [6] for reconstructing a function from x-rays from a limited range of view.

3. Formulas. The set of sources for the divergent beam x-ray transform is taken to be a sphere S_r with center at the origin and radius $r>0$. If B_r is the open ball in R^n with center at the origin and radius r, then we assume that $\overline{\Omega} \subset B_r$.

It is convenient to introduce the line integral transform

$$\mathcal{L}_a f(y) = \mathcal{L}f(a,y) = \int_{-\infty}^{\infty} f\left(a + \frac{ty}{|y|}\right) dt. \tag{3.1}$$

The following relations follow immediately from (1.1) and (2.1):

$$\mathcal{L}_a f(y) = \mathcal{D}_a f\left(\frac{y}{|y|}\right) + \mathcal{D}_a f\left(\frac{-y}{|y|}\right), \tag{3.2}$$

$$\mathcal{L}_a f(\theta) = \mathcal{P}_\theta f(E_\theta a), \tag{3.3}$$

where E_θ is the orthogonal projection in R^n on $\theta^\perp$.

Several formulas will be simpler to state with the introduction of the operator

$$\tilde{\mathcal{L}}_a f(y)=\tilde{\mathcal{L}} f(a,y)=\mathcal{L}_a f(y)|\langle a,y\rangle||y|^{-n}. \tag{3.4}$$

By the assumption on Ω and the source set either $\mathcal{D}_a f(\theta)=0$ or $\mathcal{D}_a f(-\theta)=0$, so here the three operators $\mathcal{D},\mathcal{L},\tilde{\mathcal{L}}$, are equivalent.

LEMMA 3.5. *Let ρ be a measurable function of one variable. If $\rho(\langle a,\xi\rangle)\tilde{\mathcal{L}}|f|(a,\theta)$ is an integrable function of a, then*

$$\int_{S_r}\rho(\langle a,\xi\rangle)\tilde{\mathcal{L}} f(a,\theta)\,da=2r\int_{R^n}\rho(\langle x,\xi\rangle)f(x)\,dx, \tag{3.6}$$

provided $\langle\theta,\xi\rangle=0$.

Proof. Let $S_r^+=\{a\in S_r:\langle a,\theta\rangle>0\}$ and write $a=x'+\langle a,\theta\rangle\theta$ so that $x'=E_\theta a$. Then $da=r\langle a,\theta\rangle^{-1}dx'$. Now (3.3), (3.4) and the fact that S_r surrounds Ω shows that the left-hand side of (3.6), with S_r replaced by S_r^+, is equal to

$$r\int_{B_r\cap\theta^\perp}\rho(\langle x',\xi\rangle)\mathcal{P}_\theta f(x')\,dx'=r\int_{\theta^\perp}\rho(\langle x',\xi\rangle)\mathcal{P}_\theta f(x')\,dx'.$$

Writing out $\mathcal{P}_\theta f$ as an integral and applying Fubini's theorem gives one half of the right-hand side of (3.6). Adding the contribution from the lower hemisphere (computed similarly) gives the desired result.

This lemma has several useful consequences, one of which is the analogue of (2.3).

LEMMA 3.7.

$$(2\pi)^{-n/2}\int_{S_r}e^{-i\langle a,\xi\rangle}\tilde{\mathcal{L}} f(a,\theta)\,da=2r\hat{f}(\xi),$$

provided $\langle\theta,\xi\rangle=0$.

Proof. Take $\rho(t)=e^{-it}$ in Lemma 3.5.

Remark 3.8. The Fourier transform formula above was first discovered in two dimensions by K. T. Smith [9] by computing the Fourier transform of both sides of (3.11) below.

Lemma 3.7, together with Semyanistyi's formula [8, p. 61] for the Fourier transform of even functions homogeneous of order $1-n$,

$$\left[|x|^{1-n}h\left(\frac{x}{|x|}\right)\right]\hat{}\,(\xi)=(2\pi)^{-n/2}\pi|\xi|^{-1}\int_{S^{n-1}\cap\xi^\perp}h(\phi)\,d\phi, \tag{3.9}$$

lead to an inversion formula for the divergent beam x-ray transform. A different derivation of the inversion formula and approximate inversion formulas more suitable for numerical implementation are given in [5], [9].

THEOREM 3.10. *Let $c(n)=\Gamma((n+1)/2)/(2(n-1)\pi^{(n+1)/2})$. Then*

$$f(x)=c(n)r^{-1}\Lambda\int_{S_r}\tilde{\mathcal{L}} f(a,x-a)\,da. \tag{3.11}$$

Proof. Let $F(x)=c(n)r^{-1}\int_{S_r}\tilde{\mathcal{L}} f(a,x-a)\,da$. Since $\tilde{\mathcal{L}} f(a,y)$ is even and homogeneous of order $1-n$ in the second variable, (3.9) gives

$$\left[\tilde{\mathcal{L}} f(a,\cdot)\right]\hat{}\,(\xi)=(2\pi)^{-n/2}\pi|\xi|^{-1}\int_{S^{n-1}\cap\xi^\perp}\tilde{\mathcal{L}} f(a,\phi)\,d\phi.$$

This, together with Lemma 3.7, gives

$$\hat{F}(\xi)=c(n)r^{-1}(2\pi)^{-n/2}\pi|\xi|^{-1}\int_{S_r}\int_{S^{n-1}\cap\xi^{\perp}}e^{-i\langle a,\xi\rangle}\tilde{\mathcal{L}}f(a,\phi)\,d\phi\,da \tag{3.12}$$

$$=c(n)r^{-1}(2\pi)^{-n/2}\pi|\xi|^{-1}\int_{S^{n-1}\cap\xi^{\perp}}\int_{S_r}e^{-i\langle a,\xi\rangle}\tilde{\mathcal{L}}f(a,\phi)\,da\,d\phi$$

$$=2c(n)\pi|\xi|^{-1}\int_{S^{n-1}\cap\xi^{\perp}}\hat{f}(\xi)\,d\phi$$

$$=2c(n)\pi|S^{n-2}||\xi|^{-1}\hat{f}(\xi)=|\xi|^{-1}\hat{f}(\xi),$$

the last equality following from (2.7). (Since $\tilde{\mathcal{L}}f(a,x-a)$ is locally integrable and of polynomial growth, so is its integral over S_r. The interchange of Fourier transform and integration in (3.12) is then justified since $\int_{S_r}\tilde{\mathcal{L}}f(a,x-a)\,da$ is a tempered distribution which acts on test functions by integration.) The result follows from (3.12) and (2.6).

4. The Helgason–Ludwig theorem. For $s\geq 0$, the Sobolev space $H^{s,0}(S_r\times S^{n-1})$ consists of the measurable functions u on $S_r\times S^{n-1}$ such that $\int_{S^{n-1}}\|u(\cdot,\theta)\|^2_{H^s(S_r)}d\theta<\infty$. See [3, §2.6] for a discussion of Sobolev spaces on manifolds. The Helgason–Ludwig theorem for the divergent beam x-ray transform is the following:

THEOREM 4.1. *The map $\tilde{\mathcal{L}}: L^2_0(\Omega)\to H^{1/2,0}(S_r\times S^{n-1})$ is continuous and has a continuous inverse. Moreover, if h is a measurable function on $S_r\times S^{n-1}$, then $h=\tilde{\mathcal{L}}f$ for some $f\in L^2_0(\Omega)$ if and only if*

i) $h\in H^{1/2,0}(S_r\times S^{n-1})$,

ii) $h(a,\theta)=0$ *if the line through a with direction θ misses Ω,*

iii) $h(a,\theta)=h(b,\theta)$ *if $E_\theta a=E_\theta b$,*

iv) *for each nonnegative integer m and each $\xi\in R^n$ and θ with $\langle\theta,\xi\rangle=0$, $q_{m,\theta}(\xi)=\int_{S_r}\langle a,\xi\rangle^m h(a,\theta)\,da$ is independent of the choice of θ and the function $q_m(\xi)=q_{m,\theta}(\xi)$, which is therefore well defined on R^n, is a homogeneous polynomial of degree $\leq m$.*

Proof. From (3.4) and (3.3)

$$|\langle a,\theta\rangle|^{-1}\tilde{\mathcal{L}}f(a,\theta)=\mathcal{P}_\theta f(E_\theta a). \tag{4.2}$$

Since $\bar{\Omega}\subset B_r$, $|\langle a,\theta\rangle|^{-1}$ is infinitely differentiable and bounded on a neighborhood of the support of $\tilde{\mathcal{L}}f$. This, together with the continuity of $\mathcal{P}$ and its inverse mentioned in Theorem 2.4, establishes the continuity of $\tilde{\mathcal{L}}$ and its inverse.

The necessity of i), ii) and iii) follows from (4.2), (3.1), and the above remarks. The necessity of iv) is established by taking $\rho(t)=t^m$ in Lemma 3.5.

To establish the sufficiency, first use iii) to define $g(\theta,x)$ on $T(S^{n-1})$ by

$$g(\theta,x)=h(a,\theta),\quad \text{where } x=E_\theta a. \tag{4.3}$$

CLAIM. *There exists an $f\in L^2_0(\Omega)$ such that*

$$\mathcal{P}f(\theta,x)=|\langle a,\theta\rangle|^{-1}g(\theta,x)=\left(r^2-|x|^2\right)^{-1/2}g(\theta,x). \tag{4.4}$$

To establish the claim, we check the conditions of Theorem 2.4. Conditions i) and ii) of Theorem 2.4 follow from (4.3), the fact that $\bar{\Omega}\subset B_r$, and conditions i) and ii) above. To verify iii) in Theorem 2.4, proceed as in the proof of Lemma 3.5 to obtain

$$2rP_{m,\theta}(\xi)=2r\int_{\theta^{\perp}}\langle x,\xi\rangle^m g(\theta,x)\left(r^2-|x|^2\right)^{-1/2}dx$$

$$=\int_{S_r}\langle a,\xi\rangle^m h(a,\theta)\,da=q_{m,\theta}(\xi).$$

So, condition iii) of Theorem 2.4 follows from iv) above. This establishes (4.4). Finally, (4.3), (3.3) and (3.4) give that $\tilde{\mathcal{L}}f(a,\theta)=h(a,\theta)$ completing the proof.

In light of Remark 2.5 the following extension of Theorem 4.1 holds. The proof is identical.

COROLLARY 4.5. *Theorem* 4.1 *is valid if* $L_0^2(\Omega)$ *is replaced by* $H_0^s(\Omega)$, $s\geq 0$, *provided that* $H^{1/2,0}(S_r\times S^{n-1})$ *is replaced by* $H^{t,0}(S_r\times S^{n-1})$ *with* $t=s+\frac{1}{2}$.

REFERENCES

[1] S. HELGASON, *The Radon transform on Euclidean spaces, compact two point homogeneous spaces and Grassman manifolds*, Acta Math., 113 (1965), pp. 153–180.

[2] _____, *The Radon Transform*, Birkhauser, Boston, 1980.

[3] L. HORMANDER, *Linear Partial Differential Operators*, Springer–Verlag, New York, 1969.

[4] G. N. HOUNSFIELD, *Computerized transverse axial scanning (tomography). I: Description of system*, Brit. J. Radiol., 46 (1973), pp. 1016–1022.

[5] J. V. LEAHY, K. T. SMITH AND D. C. SOLMON, *Uniqueness, nonuniqueness and inversion in the x-ray and Radon problems*, Proc. International Symposium on Ill-Posed Problems, Newark, DE Oct. 1979.

[6] A. K. LOUIS, *Picture reconstruction from projections in restricted range*, Math. Meth. Appl. Sci., 2 (1980), pp. 209–220.

[7] D. LUDWIG, *The Radon transform on Euclidean space*, Comm. Pure Appl. Math., 69 (1966), pp. 49–81.

[8] V. I. SEMYANISTYI, *Homogeneous functions and some problems of integral geometry in spaces of constant curvature*, Soviet Math. Dokl., 2 (1961), pp. 59–62.

[9] K. T. SMITH, *Reconstruction formulas in computed tomography*, Proc. AMS Short Course on Computed Tomography, Jan. 1982.

[10] K. T. SMITH, D. C. SOLMON AND S. L. WAGNER, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. AMS, 83, (1977), pp. 1227–1270.

[11] D. C. SOLMON, *The x-ray transform*, J. Math. Anal. Appl., 56 (1976), pp. 61–83.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0010 $01.25/0

ASYMPTOTIC INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS*

PHILIP HARTMAN†

Abstract. This paper concerns asymptotic integration of nth order linear differential equations which are perturbations of equations having constant coefficients and which have zero as a characteristic number. When the unperturbed equation has no other purely imaginary characteristic number, Dunkel's general result has been successively improved by Hartman and Wintner and by Prevatt. A generalization of Prevatt's result is obtained when other purely imaginary characteristic numbers are present. It is also observed that results on integration by Laplace–Stieltjes transforms suggest some questions about asymptotic integration.

1. Introduction. Let $Q(\lambda)$ be a polynomial of degree n with constant complex coefficients,

$$Q(\lambda)=\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_0, \qquad a_0\neq 0, \quad n\geq 0. \tag{1.1}$$

Consider the linear differential equation

$$Q(D)D^d u=\sum_{j=0}^{n+d-1} f_j(t)D^j u, \quad \text{where } D=\frac{d}{dt}, \quad d>0, \tag{1.2}$$

and the coefficients $f_j(t)$ are continuous complex-valued functions for large t. Suppose that the multiplicity of any purely imaginary zero of $Q(\lambda)\lambda^d$ does not exceed h ($\geq d$). It follows from a theorem of Dunkel [2] (cf. [5, Chap. XI, pp. 304, p. 321]) that if

$$\int^\infty t^{h-1}\sum_{j=0}^{n+d-1}|f_j(t)|\,dt<\infty, \tag{1.3}$$

then, for fixed $k=0,\cdots,d-1$, (1.2) has a solution satisfying $u\sim t^k$ as $t\to\infty$. When $Q(\lambda)$ has no purely imaginary root (in particular, $h=d$), Hartman and Wintner [7] (cf. [5, Thm. 17.3, p. 316]) have shown that even if the assumption (1.3) is relaxed to

$$\int^\infty\left\{\sum_{j=0}^{d-1} t^{d-1-j+\alpha}|f_j(t)|+\sum_{j=d}^{n+d-1} t^\alpha|f_j(t)|\right\}dt<\infty \tag{1.4}$$

for some $\alpha\geq 0$, then, for fixed $k=0,\cdots,d-1$, (1.2) has a solution satisfying

$$\begin{aligned} &t^{j-k+\alpha}D^j(u-t^k)\in L_0^\infty \quad \text{for } 0\leq j<d,\\ &t^{d-k-1+\alpha}D^j u\in L^1\cap L_0^\infty \quad \text{for } d\leq j<n+d. \end{aligned} \tag{1.5}$$

More recently Prevatt [9] has shown that, in this assertion, (1.4) can be further relaxed to

$$\begin{aligned} &\int^\infty\left\{\sum_{j=0}^{k} t^{d-1-j+\alpha}|f_j(t)|+\sum_{j=k+1}^{d-1} t^{d-1-j}|f_j(t)|\right\}dt<\infty,\\ &\limsup_{t\to\infty}\int_t^{t+1}\sum_{j=d}^{n+d-1}|f_j(s)|\,ds<\varepsilon_0, \end{aligned} \tag{1.6}$$

where $\varepsilon_0>0$ is a number depending only on $Q(\lambda)$.

* Received by the editors February 26, 1982, and in revised form July 20, 1982.
† 201 Glenwood Circle, Monterey, California.

The object of this paper is twofold. *The first objective is to generalize Prevatt's result to the case when* $Q(\lambda)$ *has some purely imaginary roots*; see Theorem 2.1 below and the remarks concerning it. *The second objective is to raise a related question* which unfortunately we shall have to leave open.

In order to state this question, we make some explanatory remarks. Prevatt's result was suggested in part by Hartman [4] where it is assumed that the coefficients f_j are representable as Laplace–Stieltjes transforms of suitable functions and it is asserted that (1.2) has solutions with similar representations. These results can also be viewed as results in the asymptotic integration of (1.2). For example, if (1.2) is the equation

$$(1.7) \qquad (D-i)^2 Du = f_0(t)u \quad \text{with } f_0(t) = t^{-1-\varepsilon}, \quad \varepsilon > 0$$

(so that $h=2$ and $d=1$), neither the results mentioned above nor those below are applicable. Nevertheless, it follows from [4] that (1.7) has a solution satisfying $u-1$, tu', $t^2u''=o(1)$ as $t\to\infty$. The open question which I should like to pose is the following: *Can parts of the results of* [1], [4] *be translated into a theory of asymptotic integration, free of the theory of Laplace–Stieltjes transforms*? I would guess that hypotheses in such a theory might be of the type

$$\int^{\infty} \sum_{j=0}^{n+d-1} \sum_{k=0}^{m} t^{\alpha+k} \left| D^k f_j(t) \right| dt < \infty$$

for suitable $\alpha=\alpha(j)$, $m=m(j)$. (The part of Prevatt [9] on asymptotic integration and the results below were motivated by this question.)

2. Notation and main result. Let $P_m(\lambda)$ be a polynomial of degree m with constant coefficients having only purely imaginary zeros,

$$(2.1) \qquad P_m(\lambda) = \lambda^d \prod_{\rho=1}^{\tau} (\lambda - i\lambda_\rho)^{k(\rho)}, \qquad 0=\lambda_0, \lambda_1, \cdots, \lambda_\tau \text{ distinct reals},$$

$$(2.2) \qquad m = d + \sum_{\rho=1}^{\tau} k(\rho), \qquad d = k(0) > 0, \quad k(\rho) > 0, \quad \tau \geq 0.$$

Let $Q(\lambda)$ be a polynomial of form (1.1) having no purely imaginary zeros,

$$(2.3) \qquad Q(i\lambda) \neq 0 \quad \text{for } -\infty < \lambda < \infty, \quad \text{degree } Q = n \geq 0.$$

We shall consider the linear differential equation

$$(2.4) \qquad Q(D)P_m(D)u = \sum_{j=0}^{n+m-1} f_j(t) D^j u + g(t),$$

in which f_j, g are complex-valued functions continuous for large t, say $t \geq T\ (\geq 1)$.

We deal with $m+1$ factorizations of P_m into polynomials as follows:

$$(2.5) \qquad P_m(\lambda) = P^{m-j}(\lambda) P_j(\lambda) \quad \text{for } j = 0, \cdots, m,$$

where $P^0(\lambda) \equiv 1$ and the polynomials

$$(2.6) \qquad P_0(\lambda), \cdots, P_m(\lambda) \text{ are linearly independent}$$

over the complex number field, $0 \leq \text{degree } P_j \leq m$. These polynomials will be enumerated so that

$$(2.7) \qquad 0 \leq r(0) \leq \cdots \leq r(m) = d,$$

where $r(j)$ is the multiplicity of $\lambda=0$ as a root of $P_j(\lambda)$. Let $h(j)$ be the maximum of the multiplicities of the roots of $P^{m-j}(\lambda)$, so that

$$0\le h(j)\le h \quad (\equiv \max(d,k(1),\cdots,k(\tau))). \tag{2.8}$$

The differential equation (2.4) can be written in the form

$$Q(D)P_m(D)u=\sum_{j=0}^{n-1}\phi_j(t)D^jP_m(D)u+\sum_{j=0}^{m-1}\psi_j(t)P_j(D)u+g(t), \tag{2.9}$$

where ϕ_j,ψ_j are linear combinations of the f_k with constant coefficients. Our main results will be stated in terms of (2.9), rather than (2.4).

Remark 1. The factorizations (2.5) can be chosen to satisfy

$$1\equiv P_0=P^0, \qquad P_m=P^m, \quad P_j \text{ divides } P_{j+1}, \quad \text{degree} P_j=j, \tag{2.10}$$

so that

$$h\equiv h(0)\ge h(1)\ge\cdots\ge h(m)=0,$$

and to satisfy

$$h-h(j)\ge r(j) \tag{2.11}$$

for $j=0,\cdots,m$. In fact, if $P_0\equiv 1$, then (2.11) holds for $j=0$. For a given k, suppose that $P_0,\cdots,P_k$ have been defined and satisfy (2.11) for $j=0,\cdots,k$. If $h-h(k)>r(k)$ or $h-h(k)\ge r(k)=d$, we can define $P_{k+1}=(\lambda-i\lambda^*)P_k$, where $\lambda=i\lambda^*$ is any root of P^{m-k}. If $h-h(k)=r(k)$ $(<d)$, then we can choose $\lambda^*=0$ (i.e., $r(k+1)=r(k)+1$ if and only if this makes $h(k+1)=h(k)-1$ (i.e., if and only if $\lambda=0$ is the only root of P^{m-k} of maximum multiplicity).

Remark 2. In particular, we can attain (2.5), (2.10) and (2.11) for $j=0,\cdots,m$ if the polynomials $1=P_0(\lambda),\cdots,P_{m-d}(\lambda)$ do not vanish for $\lambda=0$ (so that $r(0)=\cdots=r(m-d)=0$) and we choose $P_{m-d+j}(\lambda)=\lambda^jP_{m-d}(\lambda)$ for $j=1,\cdots,d$ (so that $r(m-d+j)=j$, $h(j)=d-r(j)$, and $h-h(j)=h-d+r(j)\ge r(j)$). In this case, the conditions on ψ_j in (2.17) in Theorem 2.1 become

$$t^{\alpha+h-1}\sum_{j=0}^{m-d-1}|\psi_j(t)|+\sum_{j=0}^{k}t^{\alpha+h-1-j}|\psi_{m-d+j}(t)|\in L^1. \tag{2.12}$$

Remark 3. In the case

$$h=d>\max(k(1),\cdots,k(\tau))\equiv\kappa, \tag{2.13}$$

we can choose $P_j(\lambda)=\lambda^j$ for $j=0,\cdots,d-\kappa$ and the other P_j suitably to satisfy (2.10) and (2.11) for $j=0,\cdots,m$. The conditions in (2.17) for $j=0,\cdots,k(<d-\kappa)$ then become $t^{\alpha+d-j-1}\psi_j\in L^1$.

Below L^p denotes the set of functions belonging to $L^p[T,\infty)$ and L_0^∞ the set of (continuous) functions in L^∞ tending to 0 as $t\to\infty$.

THEOREM 2.1. *In addition to the assumptions of continuity and notation* (2.1)–(2.8) *above, let k be an integer,*

$$0\le k\le d-1, \quad \alpha\ge 0, \quad c \text{ an arbitrary constant}, \tag{2.14}$$

$$\limsup_{t\to\infty}\int_t^{t+1}|\phi_j(s)|\,ds<\varepsilon_0 \quad \text{for } 0\le j<n, \tag{2.15}$$

$$t^{h(j)-1}\psi_j\in L^1 \quad \text{for } 0\le j<m, \tag{2.16}$$

$$t^{\alpha+h-r(j)-1}\psi_j\in L^1 \quad \text{if } 0\le j\le m \text{ and } r(j)\le k, \tag{2.17}$$

$$t^{\alpha+h-k-1}g(t)\in L^1, \tag{2.18}$$

where ε_0 *in* (2.15) *is a number depending only on the polynomial* $Q(\lambda)$. *Then* (2.9) *has a solution satisfying*

$$t^{\alpha+h-h(j)-k}P_j(D)[u-ct^k]\in L_0^\infty \quad \text{for } 0\le j<m, \tag{2.19}$$

$$t^{\alpha+h-h(j)-k-1}P_j(D)[u-ct^k]\in L^1 \quad \text{if } 0\le j<m \text{ and } h(j)\le\alpha+h-k, \tag{2.20}$$

$$t^{\alpha+h-k-1}D^jP_m(D)u\in L_0^\infty\cap L^1 \quad \text{for } 0\le j<n. \tag{2.21}$$

Remark 4. The applicability of Theorem 2.1 may depend on the choice of the factorizations (2.5).

Remark 5. The conditions (2.15), involving ε_0, are vacuous if $n=0$ (i.e., $Q(\lambda)=1$). If $n>0$, the existence of a suitable ε_0 is deduced below from a theorem of Schäffer [10] on the persistence of exponential dichotomies under certain perturbations; cf. Step (c) in the proof of Theorem 2.1 below.

COROLLARY 2.1. *When the factorizations* (2.5) *in Theorem* 2.1 *satisfy* (2.11) *for a given* j, $0\le j\le m$, *then the corresponding assumption in* (2.16) *is implied by*

$$t^{h-r(j)-1}\psi_j(t)\in L^1 \tag{2.22}$$

(*which is redundant if* $r(j)\le k$ *by* (2.17)), *and* (2.19), (2.20) *imply that*

$$\begin{aligned} &t^{\alpha+r(j)-k}P_j(D)[u-ct^k]\in L_0^\infty,\\ &t^{\alpha+r(j)-k-1}P_j(D)[u-ct^k]\in L^1 \quad \text{if } 0\le j\le m \text{ and } h(j)\le\alpha+h-k \end{aligned} \tag{2.23}$$

for the given j.

In some cases, it may be more convenient to state results in terms of (2.4), rather than its equivalent (2.9). Such a result is the following.

COROLLARY 2.2. *In* (2.1) *and* (2.4), *assume that* (2.13) *holds with* $\kappa\ge 1$ *and that the coefficient functions* g *and* f_j *satisfy* (2.18) *and*

$$t^{\alpha+d-j-1}f_j\in L^1 \quad \text{for } 0\le j\le k \quad (<d-\kappa), \tag{2.24}$$

$$t^{d-j-1}f_j\in L^1 \quad \text{for } k<j<d-\kappa, \tag{2.25}$$

$$t^{k-1}f_j\in L^1 \quad \text{for } d-\kappa\le j<m+n. \tag{2.26}$$

Then, for a fixed integer k, $0\le k<d-\kappa$, *and a constant* c, (2.4) *has a solution satisfying*

$$t^{\alpha+j-k}D^j[u-ct^k]\in L_0^\infty \quad \text{for } 0\le j<d-\kappa, \tag{2.27}$$

$$t^{\alpha+d-\kappa-k}D^ju\in L_0^\infty \quad \text{for } d-\kappa\le j<n, \tag{2.28}$$

$$t^{\alpha+d-\kappa-k-1}D^ju\in L^1 \quad \text{for } d-\kappa\le j<n. \tag{2.29}$$

The proof of Theorem 2.1 below uses the contraction principle, so that the solution $u(t)$ can be obtained by suitable successive approximations. By using Tikhonov's fixed point theorem, it is possible to obtain an analogue of Theorem 2.1 for a nonlinear equation:

$$Q(D)P_m(D)u=\Psi\big(t,u,P_1(D)u,\cdots,P_m(D)u,DP_m(D)u,\cdots,D^{n-1}P_m(D)u\big). \tag{2.30}$$

THEOREM 2.2. *In* (2.30). *let* $\Psi(t,z)$, $z=(z_0,\cdots,z_{n+m-1})$, *be continuous for large* t *and arbitrary* z *satisfying*

$$|\Psi(t,z)|\le\sum_{j=0}^{n-1}\phi_j(t)|z_{j+m}|+\sum_{j=0}^{n-1}\psi_j(t)|z_j|+g(t), \tag{2.31}$$

where $\phi_j \geq 0$, $\psi_j \geq 0$, $g \geq 0$ *are continuous for large* t *and satisfy* (2.15)–(2.18). *Then* (2.30) *has a solution* $u = u(t)$ *for large* t *satisfying* (2.19)–(2.21).

The proof of Theorem 2.2 will be omitted. It follows a variant of the proof of Theorem 2.1 similar to those in [6] and [9]; cf. [5, pp. 441–447]. Tikhonov's theorem can be replaced in this proof by the contraction principle if, in addition to (2.31), one assumes a Lipschitz condition of the form

$$|\Psi(t,x)-\Psi(t,z)| \leq \sum_{j=0}^{n-1} \phi_j(t)|x_{j+m}-z_{j+m}| + \sum_{j=0}^{m-1} \psi_j(t)|x_j-z_j|. \tag{2.32}$$

3. Proof of Theorem 2.1. We shall assume $n>0$. It will be clear that the proof can be simplified if $n=0$. For in this case, Step (c) involving exponential dichotomies can be eliminated and Step (d) only requires $w \in B_2$ rather than $w \in B_3$.

In the differential equation (2.9), let

$$u = v + ct^k \tag{3.1}$$

to obtain

$$Q(D)P_m(D)v = \sum_{j=0}^{n-1} \phi_j(t) D^j P_m(D)v + F(t), \tag{3.2}$$

where $F(t) = \Phi(t, v(t))$,

(3.3)

$$\Phi(t, v, P_1(D)v, \cdots, P_m(D)v) = \sum_{j=0}^{m-1} \psi_j(t) P_j(D)v + c \sum_{r(j)\leq k} \psi_j(t)\left[P_j(D)t^k\right] + g(t).$$

We rewrite (3.2) as a system of a pair of equations

$$Q(D)w = \sum_{j=0}^{n-1} \phi_j(t) D^j w + F(t), \tag{3.4}$$

$$P_m(D)v = w. \tag{3.5}$$

(a) *The spaces* B_1, B_2, B_3. Let B_1 be the Banach space of functions $v(t) \in C^{m-1}[T, \infty)$ with norm

$$\|v\|_1 = \sum_{j=0}^{m-1} \left\| t^{\alpha+h-h(j)-k} P_j(D)v \right\|_{L_0^\infty} < \infty. \tag{3.6}$$

Let B_2 be the Banach space of functions $F(t)$ such that $t^{\alpha+h-k-1}F(t) \in L^1$ with norm

$$\|F\|_2 = \|t^{\alpha+h-k-1}F\|_{L^1}. \tag{3.7}$$

Finally, let $B_3 \subset B_2$ be the Banach space of functions $w(t) \in C^{n-1}[T, \infty)$ with norm

$$\|w\|_3 = \sum_{j=0}^{n-1} \|t^{\alpha+h-k-1} D^j w\|_{L_0^\infty \cap L^1}. \tag{3.8}$$

(b) *The affine map* $T_1: B_1 \to B_2$, where $v \mapsto F(t) = \Phi(t, v(t))$. For $v \in B_1$, the terms $t^{\alpha+h-k-1}\psi_j(t)P_j(D)v$ of $t^{\alpha+h-k-1}F(t)$ in (3.3) are in L^1 by virtue of (3.6) and (2.16). Correspondingly, the terms $t^{\alpha+h-k-1}\psi_j(t)[P_j(D)t^k]$ are in L^1, by (2.17) and

$t^{k-r(j)}[P_j(D)t^k]\in L^\infty$. Thus $F\in B_2$ when $v\in B_1$. It is also clear from (3.3) that for $v_1, v_2\in B_1$,

$$\|T_1v_1-T_1v_2\|_2\le C_{12}\|v_1-v_2\|_1, \tag{3.9}$$

where

$$C_{12}=\int_T^\infty \sum_{j=0}^{m-1} t^{h(j)-1}|\psi_j(t)|\,dt<\infty. \tag{3.10}$$

(c) *The linear operator* $T_2: B_2\to B_3$. Since $n>0$ and none of the zeros of $Q(\lambda)$ is purely imaginary, it follows that we have an exponential dichotomy for the equation $Q(D)w=F(t)$ on any half-line $t\ge T$ in the sense of [3]; (cf. [5, pp. 478–484]). It follows from results of Schäffer [10] (cf. [8, 72A and 72C]) that there exists an $\varepsilon_0>0$ with the property that if (2.15) holds, then we also have an exponential dichotomy for (3.4). From this, it follows that there is a constant $C_{23}=C_{23}(\gamma)$ for $\gamma\ge 0$ such that if $t^\gamma F\in L^1[T,\infty)$, then (3.4) has a unique solution $w=w(t)$ satisfying $w(t_0)=0$ for a fixed t_0, $t^\gamma D^j w(t)\in L_0^\infty\cap L^1$ for $j=0,\cdots,n-1$, and that

$$\|w\|_3\le C_{23}\|F\|_2 \tag{3.11}$$

if $\gamma=\alpha+h-k-1$. The correspondence $F\mapsto w$ defines the bounded linear operator $T_2: B_2\to B_3$.

(d) *The linear operator* $T_3: B_3\to B_1$. A basis for the set of solutions of $P_m(D)v=0$ are the functions $t^j\exp(i\lambda_\sigma t)$ for $j=0,\cdots,k(\sigma)-1$ and $\sigma=0,\cdots,\tau$, where $\lambda_0=0$ and $k(0)=d$. Hence there is a unique set of constants $c_{j\sigma}^m$ such that

$$G^m(t)=\sum_{\sigma=0}^{\tau}\sum_{j=0}^{k(\sigma)-1} c_{j\sigma}^m t^j\exp(i\lambda_\sigma t) \tag{3.12}$$

is the unique solution of $P_m(D)v=0$ satisfying

$$v=Dv=\cdots=D^{m-2}v=0 \quad\text{and}\quad D^{m-1}v=1 \quad\text{at } t=0. \tag{3.13}$$

For a given $\gamma\ge 0$, write $G^m=G_{1\gamma}^m+G_{2\gamma}^m$, where

$$\begin{aligned} G_{1\gamma}^m(t)&=\sum_{\sigma=0}^{\tau}\sum_{\gamma<j<k(\sigma)} c_{j\sigma}^m t^j\exp(i\lambda_\sigma t),\\ G_{2\gamma}^m(t)&=\sum_{\sigma=0}^{\tau}\sum_{0\le j\le \min([\gamma],k(\sigma)-1)} c_{j\sigma}^m t^j\exp(i\lambda_0 t). \end{aligned} \tag{3.14}$$

It then follows that if $t^\gamma w\in L^1$, then

$$v=\int_T^t G_{1\gamma}^m(t-s)w(s)\,ds-\int_t^\infty G_{2\gamma}^m(t-s)w(s)\,ds \tag{3.15}$$

is a solution of (3.5). In particular,

$$\gamma\ge h-1\Rightarrow G_{1\gamma}^m(t)\equiv 0,\quad G_{2\gamma}^m\equiv G^m, \tag{3.16}$$

and (3.15) reduces to

$$v=-\int_t^\infty G_{2\gamma}^m(t-s)w(s)\,ds. \tag{3.17}$$

Note that, for suitable C and $t \geq T(\geq 1)$,

$$|G_{1\gamma}^m(t)| \leq Ct^{h-1} \quad \text{and} \quad |G_{2\gamma}^m(t)| \leq Ct^{\min([\gamma],h-1)}, \tag{3.18}$$

where $[\gamma]$ is the largest integer not exceeding γ.

If v is a solution of (3.5), then $y = P_j(D)v$ is a solution of $P^{m-j}(D)y = w$. Corresponding to (3.15), (3.16) and (3.18), we have (for example, by applying $P_j(D)$ to (3.15))

$$P_j(D)v = \int_T^t G_{1\gamma}^{m-j}(t-s)w(s)\,ds - \int_t^\infty G_{2\gamma}^{m-j}(t-s)w(s)\,ds, \tag{3.19}$$

$$\gamma \geq h(j)-1 \Rightarrow G_{1\gamma}^{m-j}(t) \equiv 0, \quad G_{2\gamma}^{m-j} \equiv G^{m-j}, \tag{3.20}$$

$$|G_{1\gamma}^{m-j}(t)| \leq Ct^{h(j)-1} \quad \text{and} \quad |G_{2\gamma}^{m-j}(t)| \leq Ct^{\min([\gamma],h(j)-1)}. \tag{3.21}$$

For those j for which $\gamma \geq h(j)-1$, it follows from (3.19) with $G_{1\gamma}^{m-j} = 0$ that, if $t^\gamma w \in L^1$, then

$$\left\| t^{\gamma - h(j)+1} P_j(D)v \right\|_{L_0^\infty} \leq C \| t^\gamma w \|_{L^1}, \tag{3.22}$$

$$\left\| t^{\gamma - h(j)} P_j(D)v \right\|_{L^1} \leq C \| t^\gamma w \|_{L^1} \tag{3.23}$$

for a suitable constant C.

For those j for which $\gamma < h(j)-1$, it follows from (3.19) and (3.21) that

$$|P_j(D)v(t)| \leq C\{t^{-(\gamma - h(j)+1)} + t^{-(\gamma-[\gamma])}\} \| s^\gamma w(s) \|_{L^1}.$$

Since the first exponent of t is positive and the second nonpositive, we see from (3.19) that $t^{\gamma-h(j)+1}P_j(D)v(t) \in L_0^\infty$, and that (3.22) holds for $j = 0, \cdots, m-1$.

Hence, if $\gamma = \alpha + h - k - 1$ in (3.22) and $w \in B_3 \subset B_2$, then (3.6) gives

$$\|v\|_1 \leq C_{31}\|w\|_2 \leq C_{31}\|w\|_3 \tag{3.24}$$

for a suitable constant C_{31}. The operator $T_3: B_3 \to B_1$ is defined by $w \mapsto v$, where v is given by (3.15).

(e) *Completion of the proof.* The map $S = T_3 \circ T_2 \circ T_1: B_1 \to B_1$ is well defined and is a contraction for large T in view of (3.9)–(3.10), (3.11) and (3.24). If $v = v(t) \in B_1$ is the unique fixed point of S, then (3.1) is a solution of (2.9) satisfying (2.19) and (2.21). Also (2.20) follows from (3.23). This completes the proof.

4. Proof of Corollary 2.2. Under the assumption (2.13), select the polynomials P_j as in Remark 3, so that $P_j = \lambda^j$ for $0 \leq j \leq d-\kappa$ and P_j has a factor $\lambda^{d-\kappa}$ for $d-\kappa \leq j \leq m$. Thus $h = d$ and $r(j) = j$, $h(j) = d-j$ for $0 \leq j \leq d-\kappa$, while $r(j) \geq d-\kappa$, $h(j) \leq \kappa < d-k$ for $d-\kappa < j \leq m$.

If we rewrite (2.4) in the form (2.9), then we see that $\psi_j = f_j$ for $0 \leq j \leq d-\kappa$ and ψ_j is a linear combination of $f_j, \cdots, f_{m+n-1}$ for $d-\kappa < j < m$. Thus (2.17) is equivalent to (2.24), and (2.25)–(2.26) imply (2.16) (i.e., (2.22)) for $d-\kappa < j \leq m$. Also ϕ_j is a linear combination of $f_{m+j}, \cdots, f_{n+m-1}$, so that (2.26) implies (2.15). Hence Theorem 2.1 is applicable.

Since $P_j(D) = D^j$ for $0 \leq j \leq d-\kappa$, (2.19) (i.e., (2.23)) implies (2.27). Also $P_j(D)$ is a linear combination of $D^{d-\kappa}, \cdots, D^j$ for $d-\kappa \leq j \leq m$, so that (2.19) implies (2.28) for $d-\kappa \leq j < m$ and (2.21) implies (2.28) for $m \leq j < n$. Similarly (2.20) gives (2.29) for $d-\kappa \leq j < m$ and (2.21) gives it for $m \leq j < n$.

REFERENCES

[1] J. D'ARCHANGELO AND P. HARTMAN, *Integration of ordinary linear differential equations by Laplace-Stieltjes transforms*, Trans. Amer. Math. Soc., 204 (1975), pp. 245–266.

[2] O. DUNKEL, *Regular singular points of a system of homogeneous linear differential equations of the first order*, Proc. Amer. Acad. Arts Sci., 38 (1902–3), pp. 341–370.

[3] P. HARTMAN, *On dichotomies for solutions of nth order linear differential equations*, Math. Ann., 147 (1963), pp. 57–100.

[4] _____, *On differential equations, Volterra equations and the functions* $J_\mu^2 + Y_\mu^2$, Amer. J. Math., 95 (1973), pp. 553–593.

[5] _____, *Ordinary Differential Equations*, S.M. Hartman, Baltimore, 1974 or Birkhäuser, Boston, 1982.

[6] P. HARTMAN AND N. ONUCHIC, *On the asymptotic integration of ordinary differential equations*, Pacific, J. Math., 13 (1963), pp. 1193–1207.

[7] P. HARTMAN AND A. WINTNER, *Asymptotic integration of linear differential equations*, Amer. J. Math., 77 (1955), pp. 692–724.

[8] J.L. MASSERA AND J. J. SCHÄFFER, *Linear Differential Equations and Function spaces*, Academic Press, New York, 1966.

[9] T. W. PREVATT, *Application of exponential dichotomies to asymptotic integration and the spectral theory of ordinary differential operators*, J. Differential Equations, 17 (1975), pp. 444–460.

[10] J. J. SCHÄFFER, *Linear differential equations and functional analysis* VIII, Math. Ann., 151 (1963), pp. 57–100.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0011 $01.25/0

THE LOCAL GEOMETRIC ASYMPTOTICS OF CONTINUUM EIGENFUNCTION EXPANSIONS II. ONE-DIMENSIONAL SYSTEMS*

S. A. FULLING†

Abstract. For the solutions of a system of ordinary differential equations of the type $\psi''(x)+[E(x)+p^2]\psi(x)=0$, where ψ is vector valued and $E(x)$ is an Hermitian matrix, we construct a phase-integral expansion, valid as $|p|\to\infty$, of the form $\psi(x)\sim A(x)v(x)$, where $v(x)$ is a unit vector and $A(x)$ is a positive number depending only on E and its derivatives *at* x. The terms in the series expansions of $A(x)$ and of the first-order differential equations determining $v(x)$ can be obtained as elements of a matrix series $N(x)$, which relates the basic solutions to their derivatives. If E vanishes at infinity, initial data consisting of a basis of eigenvectors of $N(x_0)$ are mapped by the phase-integral approximation into an orthogonal set of vectors at $x=\infty$. It follows that the spectral densities (normalization factors) in the eigenfunction expansion associated with the operator $-d^2/dx^2-E(x)$ and the point x_0, and hence the corresponding heat-kernel series, can be simply expressed in terms of $N(x_0)$, in analogy to the scalar results reported in [this Journal, 13 (1982), pp. 891–912]. The recursive calculation of the numerical coefficients in all these series can be efficiently computerized.

1. A vectorial WKB expansion with local amplitude. Fröman [5] has shown that the best higher-order extension of the WKB (or phase-integral) approximation of the oscillatory solutions of a second-order ordinary differential equation of Schrödinger type is a form in which both the amplitude (modulus) and the phase (argument) are taken to be power series in the expansion parameter. The amplitude function then turns out to be a *local* functional of the potential function, as opposed to an expression involving integrals ("secular terms"), and the phase function is proportional to the indefinite integral of a power (-2) of the amplitude. (We consider here only regimes of the independent variable and/or the expansion parameter which do not involve turning points.) This feature entails a number of virtues: (1) The absence of secular terms implies slower growth of the error in the approximation as a function of the interval length. (2) The identification of the derivative of the phase as the natural frequency of oscillation of the solutions near a given value of the independent variable (interpreted as time) has played an important *physical* role in the quantization of fields in a time-dependent background geometry [15], [8 and references therein]. (3) In Paper I of this series [9], the WKB–Fröman expansion was shown to be a natural tool for investigating the high-frequency asymptotic behavior of the spectral densities (normalization factors) of the Titchmarsh–Kodaira eigenfunction expansion [see [9] for references]. (This also has an intended application to general-relativistic quantum field theory, this time with a spatial interpretation of the independent variable.) Indeed, the spectral densities were seen to be intimately related to the WKB–Fröman amplitude function.

The present paper extends the conclusions of [9] to selfadjoint operators on vector-valued functions—in other words, eigenvalue equations of the form

$$H\psi(x)\equiv-\psi''(x)-E(x)\psi(x)=p^2\psi(x), \tag{1.1}$$

* Received by the editors July 23, 1981, and in revised form June 1, 1982. This research was supported in part by the National Science Foundation under grants PHY79-15229 to the Texas A&M Research Foundation and PHY77-27084 to the Institute for Theoretical Physics.

† Institute for Theoretical Physics, University of California, Santa Barbara, California 93106. Permanent address: Department of Mathematics, Texas A&M University, College Station, Texas 77843.

where ψ belongs to the space $C^\infty(\mathbb{R};\mathbb{C}^r)$ of smooth functions taking values in $\mathbb{C}^r$, and $E(x)$ is an Hermitian $r\times r$ matrix. For this purpose a generalization of the WKB–Fröman expansion to ordinary differential systems (1.1) is needed. Presumably such an expansion has many other potential applications. Some preliminary work in this direction was published earlier [7], but the problem (1.1) is slightly different from the one treated there (see Remark 1.2).

The expansion will be developed through two complementary formal approaches, which eventually will merge into a rigorous asymptotic theory.

1.1. The logarithmic-derivative matrix. We begin with a formal ansatz: Assume that there is a matrix-valued function $N(x)$ (dependent on p) such that

$$\psi'(x)=ipN(x)\psi(x) \tag{1.2}$$

for all ψ belonging to a certain space of solutions of (1.1), and that N has the asymptotic expansion

$$N\sim\sum_{s=0}^{\infty}p^{-s}N_s. \tag{1.3}$$

The scalar counterpart of (1.2)–(1.3) is the assumption that ψ can be written as

$$\psi(x)\sim\exp\left[ip\int_{x_0}^{x}\sum_{s=0}^{\infty}p^{-s}N_s(x')\,dx'\right]\qquad(r=1) \tag{1.4}$$

(cf. [5, Eq. (2)]). By analogy with the scalar case, one anticipates that half (in the sense of linear independence) of the solutions of (1.1) will belong to the space characterized by (1.2), say with $p>0$, while the others belong to the other sign of p.

Substitution of (1.2) into (1.1) yields

$$p^2(1-N^2)+ipN'+E=0, \tag{1.5}$$

or, with insertion of (1.3),

$$1-\sum_{s=0}^{\infty}\sum_{t=0}^{\infty}p^{-s}N_tN_{s-t}+i\sum_{s=1}^{\infty}p^{-s}N'_{s-1}+p^{-2}E=0. \tag{1.6}$$

Setting to zero the coefficient of each power of p we obtain recursion relations which define the formal series (1.3).

The first three recursion relations are

$$N_0^2=1\qquad\text{(the identity matrix)}, \tag{1.7a}$$

$$N_0N_1+N_1N_0=iN_0', \tag{1.7b}$$

$$N_0N_2+N_2N_0=iN_1'-N_1^2+E. \tag{1.7c}$$

There is a no loss of generality in satisfying (1.7a) by

$$N_0=1. \tag{1.8a}$$

(Choosing N_0 to be some other square root of the identity would merely associate with p some solutions of (1.1) which are more naturally associated with $-p$.) We then have from (1.7b,c)

$$N_1=0, \tag{1.8b}$$

$$N_2=\tfrac{1}{2}E. \tag{1.8c}$$

The higher-order recursion relations can now be written as

$$N_s=\frac{i}{2}N'_{s-1}-\frac{1}{2}\sum_{t=2}^{s-2}N_tN_{s-t} \qquad (s\geq 3). \tag{1.7d}$$

One obtains

$$N_3=\frac{i}{4}E', \tag{1.8d}$$

$$N_4=-\frac{1}{8}(E''+E^2), \tag{1.8e}$$

$$N_5=-\frac{i}{16}(E^{(3)}+2EE'+2E'E), \tag{1.8f}$$

$$N_6=\frac{1}{32}\left[E^{(4)}+3EE''+3E''E+5(E')^2+2E^3\right], \tag{1.8g}$$

and so on (see Tables 2 and 3).

Now recall what happens [5] in the scalar case ($r=1$). The N_s for even $s\equiv 2n$ reduce to the quantities Y_s described in [9, Thm. 4.1]. Moreover, (1.2) has the closed-form solution (1.4). Finally, the N_s for odd s are derivatives of local functionals of E:

$$iN_{2n+1}=-\tfrac{1}{2}(\ln Y)|'_{2n} \qquad (r=1), \tag{1.9}$$

where the $(\ln Y)|_{2n}$ are defined by

$$\ln\left[\sum_{n=0}^{\infty}p^{-2n}Y_{2n}(x)\right]\sim\sum_{n=0}^{\infty}p^{-2n}(\ln Y)|_{2n}(x) \tag{1.10}$$

(cf. [5, Eqs. (5)]). This allows (1.4) to be rearranged into the form

$$\psi(x)\sim Y(x)^{-1/2}\exp\left[ip\int_{x_0}^{x}Y(x')\,dx'\right], \tag{1.11}$$

$$Y\equiv\sum_{n=0}^{\infty}p^{-2n}Y_{2n} \qquad (r=1)$$

(cf. [9, Eq. (4.4)]). Thus the expansion factors into a *phase factor* of modulus 1 and an *amplitude* which is a *local* functional of E; this is the property which we wish to generalize to the vector case.

1.2. The polar decomposition. That goal is reached most easily by starting over and developing the expansion in a different way. Let $\langle\cdot,\cdot\rangle$ denote the standard inner product on $\mathbb{C}^r$, taken to be linear in the right-hand variable:

$$\langle\psi,\phi\rangle=\sum_{\alpha=1}^{r}\bar{\psi}^{\alpha}\phi^{\alpha}. \tag{1.12}$$

The following are easily verified [7, §2]:

LEMMA 1.1. *Let $\psi(x)$ and $\phi(x)$ be solutions of* (1.1) *with p^2 real and E selfadjoint. Then the generalized Wronskian*

$$\langle\psi,\phi'\rangle-\langle\psi',\phi\rangle \tag{1.13}$$

is independent of x.

LEMMA 1.2. *Any nonvanishing $\mathbb{C}^r$-valued function has the decomposition*

$$\psi(x)=A(x)u(x)e^{iS(x)}, \tag{1.14}$$

where $u(x)$ is a vector satisfying

$$\langle u,u\rangle=1, \qquad \langle u,u'\rangle=0 \tag{1.15}$$

for all x, and $A(x)>0$ and $S(x)$ are real.

COROLLARY 1.1. *For ψ given by* (1.14), *one has*

$$\langle \psi,\psi'\rangle-\langle \psi',\psi\rangle=2iS'A^2.$$

Therefore, if ψ solves (1.1), *S' is proportional to A^{-2}.*

Taking

$$S(x)=p\int_{x_0}^{x}A(x')^{-2}dx' \tag{1.16}$$

in (1.14) and substituting into (1.1), one obtains the equation

$$(1-A^{-4})u+2ip^{-1}A^{-2}u'+p^{-2}\left[\left(E+\frac{A''}{A}\right)u+2\frac{A'}{A}u'+u''\right]=0, \tag{1.17}$$

which, together with (1.15), is equivalent to (1.1). If u is a unit vector in $\mathbb{C}^r$, then $u\otimes u^*$ is the orthogonal projection operator onto u, and

$$Q\equiv 1-u\otimes u^* \tag{1.18}$$

is the projection onto the orthogonal complement of u. We split (1.17) into components parallel and perpendicular to u:

$$A^2-A^{-2}+p^{-2}\left[\langle E\rangle A^2+AA''+A^2\langle u,u''\rangle\right]=0, \tag{1.19}$$

$$u'=ip^{-1}\left[AA'u'+\tfrac{1}{2}A^2Q(Eu+u'')\right]. \tag{1.20}$$

In (1.19) the notation $\langle M\rangle$ is introduced for $\langle u,Mu\rangle$, for any matrix-valued function M. So far, no approximation has been made.

Remark 1.1. If one sets $A=B^{\nu}$, (1.17) becomes

$$(1-B^{-4\nu})u+2ip^{-1}B^{-2\nu}u'+p^{-2}\left[E+\nu(\nu-1)\left(\frac{B'}{B}\right)^2+\nu\frac{B''}{B}\right]u$$
$$+2p^{-2}\nu\left(\frac{B'}{B}\right)u'+p^{-2}u''=0.$$

Any nonzero power of A may be expanded as an asymptotic series like (1.21), with entirely equivalent results. The choice of ν is a matter of convenience. If $\nu=-\frac{1}{2}$, the expansion reduces when $r=1$ to the formalism of [5] and [1] (with $B=Y$). The choice $\nu=-\frac{1}{4}$ is most closely related to the treatment in [7].

Now introduce the ansatz

$$A(x)\sim\sum_{s=0}^{\infty}p^{-s}A_s(x), \tag{1.21}$$

$$u'(x)\sim\sum_{s=0}^{\infty}p^{-s}u'|_s(x). \tag{1.22}$$

In deriving recursion relations for the A_s and $u'|_s$, u itself (undifferentiated) is to be treated as formally independent of p. Truncated at any finite order, (1.22) will become a nonlinear differential equation for an approximate $u(x)$ which satisfies (1.15) exactly. This is the closest analogue for vector functions of the representation of a scalar function as the exponential of a power series, as in (1.4). Since the A_s and $u'|_s$ generally

depend upon u, each term of the derivatives of the series (1.21) and (1.22) will itself be a power series. This makes the calculations cumbersome. Since a tremendously more efficient algorithm will emerge in §1.3, only enough steps will be carried out here to demonstrate how (1.19)–(1.22) do determine a series in principle.

When (1.21)–(1.22) are substituted into (1.19)–(1.20), the zeroth-order equations yield immediately

$$A_0=1, \qquad u'|_0=0, \tag{1.23}$$

and the equations of order p^{-1} and p^{-2} then become

$$A_1=0, \qquad A_2=-\frac{1}{4}\langle E\rangle, \tag{1.24}$$

$$u'|_1=\frac{i}{2}QEu, \tag{1.25}$$

and

$$u'\Big|_2=\frac{i}{2}Qu''\Big|_1=\frac{i}{2}Q\big[u'|_0'|_1+u'|_1'|_0\big]=\frac{i}{2}Qu'|_1'|_0. \tag{1.26}$$

What is $u'|_1'|_0$? From (1.25) and (1.18) one obtains

$$u'|_1'=\frac{i}{2}(QE'u+QEu'-u'\langle u,Eu\rangle-u\langle u',Eu\rangle),$$

hence a sequence of equations of which the first two are

$$u'|_1'|_0=\frac{i}{2}QE'u, \tag{1.27}$$

$$\begin{aligned} u'|_1'|_1&=\frac{i}{2}\big(QEu'|_1-u'|_1\langle u,Eu\rangle-u\langle u'|_1,Eu\rangle\big)\\ &=-\frac{1}{4}(QEQEu-QEu\otimes u^*Eu+u\otimes u^*EQEu)\\ &=-\frac{1}{4}EQEu+\frac{1}{4}Q(E^2-EQE)u. \end{aligned} \tag{1.28}$$

(The second step in (1.28) uses (1.25); the third step uses (1.18). By systematically eliminating $u\otimes u^*$ in favor of Q, one obtains formulas for A_s which obviously reduce to the corresponding expressions from the scalar theory (where $A=Y^{-1/2}$) when $Q=0$.) Now (1.27), (1.26), and (1.19) yield

$$u'|_2=-\frac{1}{4}QE'u, \qquad A_3=0, \tag{1.29}$$

and (1.28) and some further calculations in the same vein lead to

$$u'|_3=-\frac{i}{8}Q(E''+E^2)u, \tag{1.30}$$

$$A_4=\frac{1}{32}\langle 2E''+5E^2-3EQE\rangle. \tag{1.31}$$

A by-product of the calculation of the A_s coefficients by (1.19) is a list of the expansion coefficients of $Y\equiv A^{-2}$:

$$Y_0=1, \qquad Y_2=\frac{1}{2}\langle E\rangle, \tag{1.32a}$$

(1.32b) $$Y_1 = Y_3 = 0,$$

(1.32c) $$Y_4 = -\frac{1}{8}\langle E'' + E^2 \rangle.$$

Remark 1.2. An expansion of the type constructed here was outlined in [7, §5] for an equation of the form

$$\psi''(x) + p^2 M(x)\psi(x) = 0.$$

(In the vector case, such an equation cannot generally be transformed to the form (1.1); contrast [9, Remark 4.2]). In [7] the matrix $M(x)$ was assumed independent of p, but the treatment can easily be generalized to a matrix with an asymptotic series

$$M \sim \sum_{s=0}^{\infty} p^{-s} M_s.$$

The present situation is then the special case

$$M_0 = 1, \quad M_2 = E, \quad M_s = 0 \quad \text{if } s \neq 0, 2.$$

1.3. Fusion of the two approaches. Since the derivative of (1.14) is

(1.33) $$\psi' = \left(ipA^{-1}u + A'u + Au'\right)e^{iS},$$

consistency of (1.2) with (1.14)–(1.16) requires

(1.34) $$u' = ipQNu,$$

(1.35) $$A^{-2} - ip^{-1}\frac{A'}{A} = \langle N \rangle.$$

Since $i^s N_s$ is selfadjoint, these equations lead to

(1.36) $$u'|_s = iQN_{s+1}u,$$

(1.37) $$A^{-2}|_s \equiv Y_s = \begin{cases} \langle N_s \rangle, & s \text{ even}, \\ 0, & s \text{ odd}, \end{cases}$$

(1.38) $$(\ln A)'|_s = \begin{cases} i\langle N_{s+1} \rangle, & s \text{ even}, \\ 0, & s \text{ odd}. \end{cases}$$

Our previous results, such as (1.32), verify (1.36) and (1.37). The expansion of A (or any other power of Y) can be obtained from that of Y by the binomial series. The relation (1.38), suitably rearranged, provides an alternative method of finding the expansion of A', or a check on the calculations; (1.38) is, of course, the r-dimensional analogue of the crucial equation (1.9) of the scalar theory. Incidentally, (1.36) and (1.37) explain the "miracle" that formulas for $u'|_s$ and Y_s never contain the projection Q in their "interiors". (Contrast (1.30) and (1.32c) with (1.31).)

THEOREM 1.1. *The differential equation*

$$\psi''(x) + \left[E(x) + \omega^2\right]\psi(x) = 0 \qquad (\psi \in C^\infty(\mathbb{R}; \mathbb{C}^r))$$

possesses, in a neighborhood of any point x_0 where the coefficient matrix E is C^∞, a basis of $2r$ solutions, each of which is asymptotic as $\omega \to +\infty$ to an expression ψ_{n_0} of the form (1.14)–(1.16) ($p = \omega$ *for* r *solutions,* $-\omega$ *for the other* r), *where A and the derivative of u are given by truncations of series of the form* (1.21) *and* (1.22). *That is, in the* 2*-norm on* $\mathbb{C}^r$,

(1.39a) $$\left\|\psi(x) - \psi_{n_0}(x)\right\| = O(\omega^{-2n_0-1}),$$

(1.39b) $$\left\|\psi'(x)-\psi'_{n_0}(x)\right\|=O(\omega^{-2n_0}),$$

uniformly on finite intervals of smoothness of E. In particular, if E is everywhere C^∞ and has compact support, then (1.39) *holds uniformly on* $\mathbb{R}$. *Every solution can be approximated by a linear combination of these $2r$ expressions. The coefficients in the expansions can be calculated by the algorithm expressed in* (1.8a, b, c), (1.7d), (1.36), (1.37). *Thus every solution is approximated by a sum of two terms, each satisfying* (1.2) *for one choice of* $\operatorname{sgn} p$.

Proof. (1) The recursion relations stemming from (1.19)–(1.22) can always be solved, and A_s is always real. (This is clear from (1.19), since

(1.40) $$\langle u,u''\rangle=-\langle u',u'\rangle$$

by (1.15).) Since $A_0=1$, when ω is sufficiently large

$$\left(\sum_{s=0}^{n_0} p^{-s}A_s\right)^{-2}$$

will be defined and asymptotically calculable by the binomial series. (If E and its derivatives are unbounded, such manipulations may not be valid uniformly in x, and one has a "turning point" problem. When $E\in C_0^\infty$, that can't happen.) Also, (1.22) (truncated) always has r linearly independent solutions for u locally, and since u is confined to the unit sphere, which is compact, these solutions extend to the entire interval of $\mathbb{R}$ considered. Thus the $2r$ expansions exist.

(2) Let $H_\omega=-d^2/dx^2-E-\omega^2$. By construction, each expression ψ_{n_0} satisfies $H_\omega\psi_{n_0}=O(\omega^{-2n_0})$ if enough terms are included in the series. Let ψ be a true solution with initial data at x_0 agreeing with ψ_{n_0} there, at least up through order $2n_0$. Then the bounds (1.39) follow as in [6, §5]: One has

$$\left\|\psi_{n_0}(x)-\psi(x)\right\|\le O(\omega^{-2n_0-1})+\int_{x_0}^{x}\left\|G_\omega(x,x')\right\|\cdot O(\omega^{-2n_0})\,dx',$$

where G_ω is the (retarded) Green matrix for the inhomogeneous system $H_\omega\psi=J$; but it is shown in [6] that

(1.41) $$\|G_\omega\|=O(\omega^{-1}),\qquad \left\|\frac{\partial G_\omega}{\partial x}\right\|=O(\omega^0).$$

In general the error terms in (1.39) grow as $|x-x_0|$ if E and its derivatives are bounded; when E, hence $H_\omega\psi_{n_0}$, has compact support, the error bound is uniform in x.

(3) Given the validity of the expansion based on (1.19)–(1.22) and the formal consistency and uniqueness of the approach based on (1.2)–(1.3), it is clear that the latter is also valid and related to the former by (1.34)–(1.38). The two final claims of the theorem follow.

Remark 1.3. As in [9, Cor. 4.1], if $E\in C_0^\infty$ the reflection and transmission coefficients vanish to all orders of the WKB approximation.

2. Inner products of basis elements. Consider now a smooth potential of compact support on $\mathbb{R}$. Let $\{u_\alpha(\infty)\}$ be an orthonormal basis for $\mathbb{C}^r$, and consider the evolution of $u_\alpha(x)$ to finite values of x under (1.22). Subscripts will be minimized by adopting the Dirac notation

(2.1) $$\langle\alpha|\beta\rangle\equiv\langle u_\alpha(x),u_\beta(x)\rangle,\qquad \langle\alpha|\beta'\rangle\equiv\langle u_\alpha(x),u'_\beta(x)\rangle,$$
$$\langle\alpha|M(x)|\beta\rangle\equiv\langle u_\alpha(x),M(x)u_\beta(x)\rangle.$$

Although $\langle\alpha|\alpha\rangle=1$ for all x by definition (1.15), it is not true that $\langle\alpha|\beta\rangle=0$ if $\alpha\neq\beta$: Solutions which start out orthogonal in $\mathbb{C}^r$ need not remain orthogonal. To see this, recall (Lemma 1.1) that $\langle\psi_\alpha,\psi_\beta'\rangle-\langle\psi_\alpha',\psi_\beta\rangle$ is a constant, which is 0 if $\langle\psi_\alpha,\psi_\beta\rangle=0$ outside the support of E. If ψ_α and ψ_β are of the form (1.14)–(1.16), one has

(2.2)
$$\langle\psi_\alpha,\psi_\beta'\rangle-\langle\psi_\alpha',\psi_\beta\rangle=\left\{\left[ip\left(\frac{A_\alpha}{A_\beta}+\frac{A_\beta}{A_\alpha}\right)+\left(A_\alpha A_\beta'-A_\alpha' A_\beta\right)\right]\langle\alpha|\beta\rangle\right.$$
$$\left.+A_\alpha A_\beta\left(\langle\alpha|\beta'\rangle-\langle\alpha'|\beta\rangle\right)\right\}e^{ip\int(A_\alpha^{-2}-A_\beta^{-2})}.$$

If $\langle\alpha|\beta\rangle$ were zero for all x, it would follow that $\langle\alpha|\beta'\rangle-\langle\alpha'|\beta\rangle=0$. On the other hand, it would also be true that $\langle\alpha|\beta\rangle'=\langle\alpha|\beta'\rangle+\langle\alpha'|\beta\rangle=0$, and hence that $\langle\alpha|\beta'\rangle=0=\langle\alpha'|\beta\rangle$. But that, of course, is generically false, since $u_\alpha'(x)$ is in general a nonzero vector orthogonal to $u_\alpha(x)$.

Nevertheless, the u_α can be chosen so that (asymptotically) $\{u_\alpha(x_0)\}$ is an orthonormal basis of eigenvectors of the Hermitian part of $N(x_0)$. The calculations which show this appear rather "formal", since the series (1.3) which would define N usually does not converge. They are to be interpreted as shorthand for order-by-order recursive relationships among the terms of asymptotic series.

One can derive a differential equation which is satisfied up to the appropriate order by the inner product of any two solutions of truncated equations (1.22). By (1.34) we have

$$\langle\alpha|\beta\rangle'=\langle\alpha|\beta'\rangle+\langle\alpha'|\beta\rangle=ip\left(\langle\alpha|Q_\beta N|\beta\rangle-\langle\alpha|N^*Q_\alpha|\beta\rangle\right),$$

from which we note in passing that $\langle\alpha|\beta\rangle'$ is of order $O(p^{-1})$ by virtue of (1.8a, b) (or (1.23)). Using (1.18), this is reduced to

$$\langle\alpha|\beta\rangle'=ip\left(\langle\alpha|N^*|\alpha\rangle-\langle\beta|N|\beta\rangle\right)\langle\alpha|\beta\rangle+ip\langle\alpha|(N-N^*)|\beta\rangle. \tag{2.3}$$

On the other hand, it is easy to show from (1.5), (1.34), and (1.18) that

$$\langle\alpha|N|\beta\rangle'=ip\langle\alpha|\beta\rangle+ip^{-1}\langle\alpha|E|\beta\rangle-ip\langle\alpha|N^*N|\beta\rangle$$
$$+ip\left(\langle\alpha|N^*|\alpha\rangle-\langle\beta|N|\beta\rangle\right)\langle\alpha|N|\beta\rangle,$$
$$\langle\alpha|N^*|\beta\rangle'=-ip\langle\alpha|\beta\rangle-ip^{-1}\langle\alpha|E|\beta\rangle+ip\langle\alpha|N^*N|\beta\rangle$$
$$+ip\left(\langle\alpha|N^*|\alpha\rangle-\langle\beta|N|\beta\rangle\right)\langle\alpha|N^*|\beta\rangle,$$

and hence the matrix element of $\hat{N}\equiv\frac{1}{2}(N+N^*)$ satisfies a homogeneous variant of (2.3):

$$\langle\alpha|\hat{N}|\beta\rangle'=ip\left(\langle\alpha|N^*|\alpha\rangle-\langle\beta|N|\beta\rangle\right)\langle\alpha|\hat{N}|\beta\rangle. \tag{2.4}$$

Outside the support of E, $N(x)$ reduces to the identity. Therefore, if $\langle\alpha|\beta\rangle=0$ at ∞, then $\langle\alpha|\hat{N}|\beta\rangle=0-$at ∞, and hence at all x because of (2.4). As a corollary (or by a parallel argument) one has

$$\langle\alpha|\beta\rangle=\langle\alpha|(1-\hat{N})|\beta\rangle \tag{2.5}$$

for all x, if $\langle\alpha|\beta\rangle=0$ at ∞.

Now let $u_\beta(x)$ be an eigenvector of $\hat{N}(x_0)$, and suppose that the eigenvalue is nonzero (as will certainly be the case if ω is large enough). Then $\langle\alpha|\beta\rangle$ at x_0 is a numerical multiple of $\langle\alpha|\hat{N}(x_0)|\beta\rangle$, and it follows from the preceding paragraph that $\langle\alpha|\beta\rangle=0$ *at* ∞ *if and only if* $\langle\alpha|\beta\rangle=0$ *at* x_0. (At intervening values of x, $\langle\alpha|\beta\rangle$ may be nonzero.) Therefore, if we choose $\{u_\alpha(x_0)\}$ to be an orthonormal basis of eigenvectors of the Hermitian matrix $\hat{N}(x_0)$ (truncated), then the corresponding vectors $\{u_\alpha(\infty)\}$ will also be orthonormal (up to the order of the approximation). By Remark 1.3 (or by symmetry) the same is true of $\{u_\alpha(-\infty)\}$.

3. The mean spectral measures. In [9], the local spectral asymptotic problem for an arbitrary selfadjoint operator that looks locally like H (1.1) was reduced to a scattering problem for H, where $E\in C_0^\infty$. In terms of continuum eigenfunctions $\phi_{\omega\alpha}$ normalized at ∞ according to [9, (4.1)], H has a spectral decomposition given by [9, (4.2)] or, equivalently (cf. [9, (4.6)]),

$$\begin{aligned}(3.1)\qquad &\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda;x_0)\psi_{\lambda k}(y)^*\\ &\equiv dE_\lambda(x,y)\\ &=\frac{1}{2\pi}\sum_{\alpha=1}^{r}\left[\phi_{\omega\alpha}(x)\otimes\phi_{\omega\alpha}(y)^*+\phi_{-\omega\alpha}(x)\otimes\phi_{-\omega\alpha}(y)^*\right]d\omega.\end{aligned}$$

(There is an implicit tensor product in the left-hand side of (3.1). Recall that $\omega=\lambda^{1/2}=|p|$.) Spectral densities are defined by

$$(3.2)\qquad d\mu^{jk}(\lambda;x_0)=\pi^{-1}\rho^{jk}(\omega;x_0)\,d\omega.$$

THEOREM 3.1. *Let H be a selfadjoint operator in an r-dimensional vector bundle over a one-dimensional manifold M, with the local representation*

$$H=-\frac{d^2}{dx^2}-E(x)$$

in a neighborhood of an interior point x_0 where E is C^∞. The spectral densities of H relative to x_0 have mean asymptotic expansions

$$(3.3)\qquad \rho^{00}(x_0)\sim\hat{N}(x_0)^{-1},$$

$$(3.4)\qquad \rho^{10}(x_0)\sim-\omega\tilde{N}(x_0)\hat{N}(x_0)^{-1},\qquad \rho^{01}\sim-\omega\hat{N}(x_0)^{-1}\tilde{N}(x_0),$$

$$(3.5)\qquad \rho^{11}(x_0)\sim\omega^2\left[\hat{N}(x_0)+\tilde{N}(x_0)\hat{N}(x_0)^{-1}\tilde{N}(x_0)\right],$$

where

$$(3.6)\qquad \begin{aligned}\hat{N}&\equiv\frac{1}{2}(N_\omega+N_\omega^*)=\sum_{s\text{ even}}\omega^{-s}N_s,\\ \tilde{N}&\equiv\frac{1}{2i}(N_\omega-N_\omega^*)=-i\sum_{s\text{ odd}}\omega^{-s}N_s,\end{aligned}$$

and N_s is defined in §1.1.

Warning to physicists. N^* is the adjoint, not the complex conjugate.

Remark 3.1. The last two sentences of [9, Thm. 4.2] apply here as well.

Proof. (1) If $M=\mathbb{R}$ and E has compact support, the form of $\phi_{\omega\alpha}$ in the limit of large ω is given in Theorem 1.1 and Remark 1.3. Thus (3.1) is

$$\begin{aligned}(3.7)\qquad \pi^{-1}\rho^{00}(x_0)\,d\omega \equiv dE_\lambda(x_0,x_0)\\ =\frac{1}{2\pi}\sum_{\alpha=1}^{r}\big[A_{\omega\alpha}(x_0)^2 u_{\omega\alpha}(x_0)\otimes u_{\omega\alpha}(x_0)^*\\ +A_{-\omega\alpha}(x_0)^2 u_{-\omega\alpha}(x_0)\otimes u_{-\omega\alpha}(x_0)^*\big]\,d\omega.\end{aligned}$$

By (1.37), $A_{\omega\alpha}^2$ is an even function of ω, and in fact $A_{\omega\alpha}^2=\langle\alpha|\hat N_\omega|\alpha\rangle^{-1}$. (Recall that the terms of N_p which are even under Hermitian conjugation are precisely those which are even in p.) The point of §2 is that $\{u_{\pm\omega\alpha}(x_0)\}$ can be chosen to be a basis of eigenvectors of $\hat N(x_0)$, consistently with the orthonormalization at ∞ required for (3.1) to be correct. Then (3.7) becomes

$$(3.8)\qquad \rho^{00}(x_0)=\sum_{\alpha=1}^{r}\big\langle\alpha|\hat N(x_0)|\alpha\big\rangle^{-1}u_{\omega\alpha}(x_0)\otimes u_{\omega\alpha}(x_0)^*,$$

which is recognized as the spectral decomposition of the matrix $\hat N(x_0)^{-1}$. Thus we arrive finally at the basis-independent formula (3.3).

(2) The other densities are obtained by differentiating (3.1) before setting $x=y=x_0$ (see [9, (3.1)]). But by (1.2) we have

$$\phi'_{p\alpha}(x)=ipN(x)\phi_{p\alpha}(x),\qquad \phi^{*\prime}_{p\alpha}(x)=-ip\phi_{p\alpha}{}^*N(x)^*.$$

In ρ^{10} and ρ^{01} the sum over $p=\pm\omega$ selects the anti-Hermitian part of N, and formulas (3.4) result. In ρ^{11} we encounter the combination

$$\begin{aligned}N_\omega\hat N^{-1}N_\omega^*+N_{-\omega}\hat N^{-1}N_{-\omega}^*&=(\hat N+i\tilde N)\hat N^{-1}(\hat N-i\tilde N)+(\hat N-i\tilde N)\hat N^{-1}(\hat N+i\tilde N)\\ &=2(\hat N+\tilde N\hat N^{-1}\tilde N),\end{aligned}$$

whence (3.5).

(3) As explained in [9], the general case can be replaced by a locally equivalent scattering operator of the type just treated. The expansions (3.3)–(3.6) are valid (unless x_0 is an endpoint of M or a point of singularity of E) in the "mean" or "effective" sense: They correctly reproduce purely locally determined entities such as the heat-kernel expansion (e.g. [11], [12]) and the short-distance divergences of Green functions (e.g., [4], [10]).

In the scalar case we have

$$\hat N=Y,\qquad \tilde N=-\frac{1}{2\omega}(Y^{-1})'/Y^{-1}$$

(see (1.35), or (1.37) and (1.9)). Thus (3.3)–(3.5) are consistent with the results [9, (4.9), (4.13), (4.14)].

4. Computational considerations. The number of terms (linearly independent monomials in the noncommuting variables $E,E',\cdots,E^{(n)},\cdots$) in the asymptotic formulas for N,ρ^{jk}, and the heat kernel

$$(4.1)\qquad K(t,x_0,x_0)\sim(4\pi t)^{-m/2}\sum_n a_n(x_0)t^n$$

increases rapidly with order. ("Order" is defined, as in [9, Thm. 3.2], to be the sum of the differential order of each factor, plus twice the number of factors; for example, N_6 (1.8g) consists of terms of order 6. In the expansion of a quantity such as N or ρ^{10}, the terms multiplying a given power of the expansion parameter are of a fixed order—e.g., a_n in (4.1) is of order $2n$. Typically the terms are all even-order or all odd-order and are reindexed accordingly.) Hand calculation soon becomes time consuming and error prone, and even listing the results in print presents practical difficulties beyond a certain point. These effects are even more noticeable for the heat-kernel expansion on a manifold of dimension $m>1$; a practical barrier has been reached at around orders 4 and 6. See Table 1.

TABLE 1

Number of linearly independent terms potentially contributing to N_{2n}, ρ_n^{00}, *or* a_n (*order* $2n$)

n	$m=1$, $r=1$,	$m=1$, $r>1$	$m>1$	$m>1$*
reference	[1],[9]	Table 2	[11]	[3]
1	1	1	2	13
2	2	2	8	~100
3	4	4	46	
4	7	9		
5	12	21		
6	21	51		
7	34	127		
8	55	322		
9	88	826		
10	137	2135		

* Background metric (used to define curvature and covariant derivatives) $\neq$ metric determined by principal symbol of operator.

For the terms of simplest structure, $E^{(s-2)}$ and $E^{s/2}$, it is easy to determine the coefficients once and for all as a function of the order, s ([1], [12]), and this could probably be done for some other classes of terms. It is clear, however, that a complete solution in closed form of the recursion relations is unlikely to be achieved, since new types of terms arise in each order, and even to devise a convenient notation, or labelling scheme, for all the possible terms is a nontrivial problem in higher dimensions, where contractions over tensor indices are involved. Furthermore, it should be remembered that recursive schemes are often more efficient in practice than evaluation of explicit combinatorial formulas; the usefulness of Pascal's triangle for the binomial coefficients is a familiar example.

Beyond the lowest orders, therefore, resort to machine calculation is inevitable. Not only is a computer needed to obtain the coefficients in the first place, but also one anticipates that in most applications of such lengthy lists of numbers, it will be better to generate them by subroutine within the computer program which uses them, rather than to copy them out of published tables.

The computation of the one-dimensional expansions, defined in the present paper, at orders $\gtrsim 8$ provides an instructive exercise preliminary to attacking the more difficult higher-dimensional problems.

Previous computations of this nature (e.g., [1], [7]) have made use of general-purpose algebraic symbol-manipulation programs such as FORMAC, TRIGMAN, and the current favorite, MACSYMA. These require entire algebraic expressions to be entered

into and processed by the computer. The author's experience [7] has been that the machine's memory capacity is exceeded within very few orders beyond the limit of feasible hand calculation. It seems more sensible to develop programs specifically adapted to these rather specialized and (once the theory of each series has been worked out) stereotyped calculations. As much as possible, the machine should manipulate lists of coefficients instead of whole expressions. Because of the recursive structure of the computation, a programming language that admits recursive subroutine calls is highly desirable.

In addition, in the project reported here an "analytic" approach (as opposed to "synthetic") was adopted: The program was written to calculate, as a self-contained problem, the coefficient of a single given monomial in the expansion of a given quantity, by enumerating all possible ways in which a term of that form could arise in the synthetic calculation of the entire expansion. For example, to calculate the coefficient of E^3 in N_6 (see (1.8g)) from (1.7d), one first finds all occurrences of E^3 in N_5' (namely, none), and second considers all ways (namely, two) in which E^3 can be writtten as a product of two terms, one of order t and one of order $6-t$; then the coefficients of those terms in N_t and N_{6-t} are found (by recursive subroutine calls) and multiplied. Each expansion is regarded as a function with the monomial as argument; a conveniently coded representation of the monomial (e.g., 0-1 for EE') is the argument passed to the subroutine. The attitude taken here was that the highest priority was to hold intermediate storage and recovery of data to an absolute minimum. This method is probably not the most efficient for producing a list of all coefficients of several successive orders, since coefficients of low order are not saved after output, but must be recomputed in finding those of higher order. However, this technique would be preferred in an application where only certain coefficients are relevant—or whenever memory is at a premium. Note also that there are tricks for calculating certain classes of coefficients by examination of special cases [12], [13]; a program of the type described can be used in a complementary way to calculate the remaining coefficients efficiently.

The computations were done on a small computer (VAX-11). The programming language used is named "C" [14]. The algorithms will now be described briefly.

The first step is to generate a list of all possible terms of a given order, encoded as described above. This was quite routine in the one-dimensional case. This input file (or its even or odd part) is passed through the programs which calculate the coefficients for $N, \rho^{00}, \rho^{10}, \rho^{01}$, and ρ^{11}.

It is convenient to introduce for the terms in the asymptotic series the notations

$$N \sim 1 + \sum_{s=2}^{\infty} p^{-s} \frac{i^{s-2}}{2^{s-1}} \mathbf{N}_s, \tag{4.2a}$$

$$\rho^{00} \sim 1 + \sum_{n=1}^{\infty} \omega^{-2n} \frac{(-1)^n}{2^{2n-1}} \boldsymbol{\rho}_n^{00}, \tag{4.2b}$$

$$\rho^{10} \sim \sum_{n=1}^{\infty} \omega^{-2n} \frac{(-1)^n}{2^{2n}} \boldsymbol{\rho}_n^{10} \tag{4.2c}$$

(and similarly for ρ^{01}),

$$\rho^{11} \sim \omega^2 + \sum_{n=1}^{\infty} \omega^{2-2n} \frac{(-1)^{n-1}}{2^{2n-1}} \boldsymbol{\rho}_n^{11}. \tag{4.2d}$$

Then the coefficient of each monomial in $\mathbf{N}_s$ or $\boldsymbol{\rho}_n^{jk}$ is a positive integer. The fundamental recursion relation (1.7d) becomes

$$\mathbf{N}_s = \mathbf{N}'_{s-1} + \sum_{t=2}^{s-2} \mathbf{N}_t \mathbf{N}_{s-t} \qquad (s>2). \tag{4.3a}$$

As described previously, the program calculates the coefficient of a given sth-order monomial in $\mathbf{N}_s$ by calculating its coefficient in each term on the right-hand side of (4.3a) and adding the results. Next, according to (3.3), we must obtain ρ^{00} as the matrix inverse of the even part of N; that is,

$$1 = \hat{N}\rho^{00} \sim 1 + \sum_{n=1}^{\infty} \omega^{-2n} \left\{ \frac{(-1)^{n-1}}{2^{2n-1}} \mathbf{N}_{2n} + \frac{(-1)^n}{2^{2n-1}} \boldsymbol{\rho}_n^{00} + \sum_{m=1}^{n-1} \frac{(-1)^{n-1}}{2^{2n-2}} \mathbf{N}_{2m} \boldsymbol{\rho}_{n-m}^{00} \right\},$$

whence

$$\boldsymbol{\rho}_n^{00} = \mathbf{N}_{2n} + 2 \sum_{m=1}^{n-1} \mathbf{N}_{2m} \boldsymbol{\rho}_{n-m}^{00} \qquad (n>0). \tag{4.3b}$$

This relation can be programmed similarly to (4.3a). (Of course, the recursion relations for $\boldsymbol{\rho}_n^{00}$ have an explicit solution in the form of a geometric series in $1-\hat{N}$, but evaluating the terms of that series in practice turns out to be virtually equivalent to working with (4.3b).) The formulas (3.4) and (3.5) for the other ρ^{jk} series require merely multiplication of known matrix series, and the order-by-order formulas have a structure quite similar to (4.3b):

$$\boldsymbol{\rho}_n^{10} = \mathbf{N}_{2n+1} + 2 \sum_{m=1}^{n-1} \mathbf{N}_{2n-2m+1} \boldsymbol{\rho}_m^{00}, \tag{4.3c}$$

$$\boldsymbol{\rho}_n^{01} = \mathbf{N}_{2n+1} + 2 \sum_{m=1}^{n-1} \boldsymbol{\rho}_m^{00} \mathbf{N}_{2n-2m+1}, \tag{4.3d}$$

$$\boldsymbol{\rho}_n^{11} = \mathbf{N}_{2n} + 2 \sum_{m=1}^{n-2} \boldsymbol{\rho}_m^{10} \mathbf{N}_{2n-2m-1}, \qquad (n>0). \tag{4.3e}$$

In evaluating the sums in (4.3b–e) it is necessary to check each possible factorization of the monomial for proper parity of the orders of the factors.

The VAX was able to compute all coefficients through order $s=14$ before time became a problem. The results through $s=10$ are presented in Tables 2 and 3. (They extend (1.8) and generalize the formulas for the commutative case in [9, §4].) Orders 11–16 and program listings are available on request.

The relation between the spectral density ρ^{00} and the heat-kernel expansion (4.1) is stated in [9, Thm. 3.1]. Combining [9, (3.11)] with (4.2b), one finds that the coefficient of each term in a_n is obtained by dividing the corresponding number in Table 2 by

$$h_n \equiv 2^{n-1}(2n-1)!!, \tag{4.4a}$$

$$h_n = 2(2n-1)h_{n-1} \tag{4.4b}$$

(Table 4). For example,

$$(4.5)\quad a_4=\frac{1}{840}\{E^{(6)}+7(EE^{(4)}+E^{(4)}E)+14(E'E^{(3)}+E^{(3)}E')$$
$$+21(E'')^2+21(E^2E''+E''E^2)+28EE''E$$
$$+28[E(E')^2+(E')^2E]+14E'EE'+35E^4\}$$

(cf. [12, Thm. 2.2]).

TABLE 2

Coefficients of terms of even order, $s=2n$. (= adjoint of term = reversed factor ordering.)*

Term	N	ρ^{00}	ρ^{11}
Order 2 (1 term):			
E	1	1	1
Order 4 (2 terms):			
E''	1	1	1
E^2	1	3	1
Order 6 (4 terms):			
$E^{(4)}$	1	1	1
$EE''+*$	3	5	3
$(E')^2$	5	5	7
E^3	2	10	2
Order 8 (9 terms):			
$E^{(6)}$	1	1	1
$EE^{(4)}+*$	5	7	5
$E'E^{(3)}+*$	14	14	16
$(E'')^2$	19	21	19
$E^2E''+*$	9	21	9
$E(E')^2+*$	18	28	22
$EE''E$	12	28	12
$E'EE'$	14	14	26
E^4	5	35	5
Order 10 (21 terms):			
$E^{(8)}$	1	1	1
$EE^{(6)}+*$	7	9	7
$E'E^{(5)}+*$	27	27	29
$E''E^{(4)}+*$	55	57	55
$(E^{(3)})^2$	69	69	71
$E^2E^{(4)}+*$	20	36	20
$EE'E^{(3)}+*$	68	96	72
$E(E'')^2+*$	102	150	102
$EE^{(3)}E'+*$	80	108	88
$EE^{(4)}E$	30	54	30
$E'EE^{(3)}+*$	48	48	64
$(E')^2E''+*$	140	150	158
$E'E''E'$	162	162	202
$E''EE''$	62	78	62
$E^3E''+*$	28	84	28
$E^2(E')^2+*$	60	126	70
$E^2E''E+*$	42	126	42
$EE'EE'+*$	56	84	84
$E(E')^2E$	76	168	84
$E'E^2E'$	42	42	98
E^5	14	126	14

TABLE 3

Coefficients of terms of odd order, $s=2n+1$. (The coefficient of $M\equiv AB$ in ρ^{10} equals that of $M^*\equiv BA$ in ρ^{01}. In N the coefficients of M and M^* are equal.)

Term	N	ρ^{10}	ρ^{01}
Order 3 (1 term):			
E'	1	1	1
Order 5 (3 terms):			
$E^{(3)}$	1	1	1
EE'	2+*	2	4
Order 7 (8 terms):			
$E^{(5)}$	1	1	1
$EE^{(3)}$	4+*	4	6
$E'E''$	9+*	11	9
E^2E'	5+*	5	15
$EE'E$	6	10	10
Order 9 (21 terms):			
$E^{(7)}$	1	1	1
$EE^{(5)}$	6+*	6	8
$E'E^{(4)}$	20+*	22	20
$E''E^{(3)}$	34+*	34	36
$E^2E^{(3)}$	14+*	14	28
$EE'E''$	38+*	42	56
$EE''E'$	42+*	42	70
$EE^{(3)}E$	20	28	28
$E'E\,E''$	28+*	42	28
$(E')^3$	60	70	70
E^3E'	14+*	14	56
$E^2E'E$	18+*	28	42

TABLE 4

Denominators of the heat-kernel expansion

n	h_n
1	1
2	6
3	60
4	840
5	15,120
6	332,640
7	8,648,640
8	259,459,200
9	8,821,612,800
10	335,221,286,400

It is not claimed that the most efficient way to calculate the ρ^{jk} is to obtain them from N (i.e., the phase-integral expansion) as has been done here. The DeWitt–Christensen–Perelomov algorithm [4], [2], [16], similarly mechanized, would give the expansions of the heat kernel and its derivatives (hence ρ^{jk}) more directly. The methods of [12], [13], when applicable, are even quicker, but it is not always easy to find such a method to determine every coefficient of interest. The method of the present paper is certainly feasible, however, and has the conceptual advantage of demonstrating the close connection between the spectral asymptotics and a phase-integral approximation of independent interest.

Acknowledgments. I thank the Institute for Theoretical Physics for its hospitality, and especially John Richardson for helping me become familiar with the computer facilities. I thank B. Bourgeois for a careful reading of the manuscript and its companions.

REFERENCES

[1] J. A. CAMPBELL, *Computation of a class of functions useful in the phase-integral approximation*. I, J. Comput. Phys., 10 (1972), pp. 308–315.

[2] 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.

[3] S. M. CHRISTENSEN AND S. A. FULLING, unpublished.

[4] B. S. DEWITT, *Dynamical Theory of Groups and Fields*, Gordon and Breach, New York, 1965.

[5] N. FRÖMAN, *Outline of a general theory for higher order approximations of the JWKB-type*, Arkiv Fysik, 32 (1966), pp. 541–548.

[6] S. A. FULLING, *Adiabatic expansions of solutions of coupled second-order linear differential equations*. I, J. Math. Phys., 16 (1975), pp. 875–883.

[7] _____, *Adiabatic expansions of solutions of coupled second-order linear differential equations*. II, J. Math. Phys., 20 (1979), pp. 1202–1209.

[8] _____, *Remarks on positive frequency and Hamiltonians in expanding universes*, Gen. Relativity Gravitation, 10 (1979), pp. 807–824.

[9] _____, *The local geometric asymptotics of continuum eigenfunction expansions*. I, this Journal, 13 (1982), pp. 891–912.

[10] 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.

[11] P. B. GILKEY, *The spectral geometry of a Riemannian manifold*, J. Differential Geom., 10 (1975), pp. 601–618.

[12] _____, *Recursion relations and the asymptotic behavior of the eigenvalues of the Laplacian*, Compositio Math., 38 (1979), pp. 201–240.

[13] _____, *The spectral geometry of the higher order Laplacian*, Duke Math. J., 47 (1980), pp. 511–528.

[14] B. W. KERNIGHAN AND D. M. RITCHIE, *The C Programming Language*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[15] L. PARKER AND S. A. FULLING, *Adiabatic regularization of the energy-momentum tensor of a quantized field in homogeneous spaces*, Phys. Rev. D, 9 (1974), pp. 341–354.

[16] A. M. PERELOMOV, *Schrödinger equation spectrum and Korteweg–de Vries type invariants*, Ann. Inst. H. Poincaré Sect. A (N.S.), 24 (1976), pp. 161–164.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0012 $01.25/0

GLOBAL PROPERTIES OF PROPER LIPSCHITZIAN MAPS*

B. H. POURCIAU†

Abstract. Equations of the form $f(x)=y$, where f is a locally Lipschitz continuous map from R^n into R^n, arise naturally in the study of nonlinear electrical networks and integrability questions in mathematical economics. In such contexts, properties of the image $f(R^n)$ become important. In this paper we use a generalized set-valued derivative and a special degree function to study the ontoness and interiority of locally Lipschitz continuous maps.

1. Introduction. Equations of the form $f(x)=y$ where $f: R^n \to R^n$ arise in the study of nonlinear resistive and dynamical electrical network equations. If f is onto, these equations are solvable for every input signal y, and if f is one-to-one, the solution signal x is unique. Since state equations for dynamical systems come from inverting resistive equations and Lipschitz continuity makes the differential equations uniquely solvable, it is natural to suppose f is locally Lipschitz continuous. Facing researchers in electrical network theory then is the problem of determining global mapping properties of locally Lipschitz continuous maps $f: R^n \to R^n$. For example, look at Chua and Lam [3], Fujisawa and Kuh [11], Haneda [13], Kawamura [15] and Kojima and Saigal [16]. The same mathematical problem is met by economists investigating integrability questions and the existence and uniqueness of equilibria in competitive economies. Read for instance Arrow and Debreu [1], Berger and Meyers [2] and Dierker [9].

In Pourciau [21], [22], [23] global properties of locally Lipschitz continuous maps are studied using a generalized set-valued derivative. The present work continuous that study.

Let us say $f: R^n \to R^n$ is *Lipschitzian* if it is locally Lipschitz continuous: each point has a neighborhood U and a constant M such that $|f(x)-f(z)| \leq M|x-z|$ for all x and z in U. By a deep theorem essentially due to Rademacher (see Federer [10]), the Fréchet derivative $f'(z)$ of a Lipschitzian map of finite-dimensional spaces exists almost everywhere (μ). Here μ stands for n-dimensional Lebesgue measure. The *generalized derivative* $\partial f(x)$ of f at x is the set

$$\bigcap_{\delta>0} \overline{\text{co}}\{f'(z): |z-x|<\delta, f'(z) \text{ exists}\}.$$

The notation $\overline{\text{co}}$ means the closure of the convex hull. For details concerning this generalized derivative and its many pleasant properties (chain rules, inverse and implicit mapping theorems, necessary conditions in optimization theory and so on) peruse Clarke [7], where the generalized derivative was first introduced, Hiriart-Urruty [14] or Pourciau [20].

A few of these properties should be recorded here. For any x, $\partial f(x)$ is a nonempty, convex, compact set of linear maps (or matrices if you prefer). As we shall be extending certain results for C^1 maps, it is crucial to us that $\partial f(x)$ reduces to the singleton $\{f'(x)\}$ whenever f is continuously differentiable on a neighborhood of x. Another result vital for the present study is the following extension of the classical (C^1) inverse function theorem. Consult Clarke [8] or Pourciau [20] for the proof. We call $\partial f(x)$ *invertible* provided it contains only invertible linear maps.

* Received by the editors September 21, 1981.
† Department of Mathematics, Lawrence University, Appleton, Wisconsin 54912.

INVERSE MAP THEOREM. *A Lipschitzian map f having an invertible generalized derivative $\partial f(x)$ at every x is a local homeomorphism.*

2. Global properties. Information about the generalized derivative will enable us to predict global mapping properties for Lipschitzian f. For continuous $f: R^n \to R^n$ it is well known (see Hadamard [12] and Palais [18]) that f is a homeomorphism if it is a proper local homeomorphism. (We say f is *proper* when $f^{-1}(K)$ is compact for all compact sets K. In finite-dimensional spaces, this is equivalent to requiring that $|f(x)| \to \infty$ if $|x| \to \infty$.) In view of our inverse map theorem, we have

THEOREM A. *A proper Lipschitzian map $f: R^n \to R^n$ is a homeomorphism if the generalized derivative $\partial f(x)$ is invertible at every x.*

In applications to electrical network equations, one often meets Lipschitzian maps whose generalized derivative is invertible off some (usually "thin") set S. Either one knows $\partial f(x)$ is *singular* (that is, contains a singular linear map) on S or one knows nothing about $\partial f(x)$ on S. Under these conditions, which global mapping properties of f remain? From Plastock [19] (see also Church and Hemmingsen [6]), f is a local homeomorphism at x if f is a local homeomorphism at each point in a punctured neighborhood of x. This hands us

THEOREM B. *A proper Lipschitzian map $f: R^n \to R^n$ is a homeomorphism if the generalized derivative $\partial f(x)$ is invertible everywhere off a discrete set S.*

In applications, unfortunately, when S is not empty it is usually not discrete either. It *is* safe to assume, though, that S has measure (μ) zero. Moreover, in applications to electrical network equations and economic equilibria the maps f often satisfy a sign condition on the determinants of the derivatives, a condition generally at least as strong as this:

CONSTANT SIGN PROPERTY. *For all x in R^n and all A in $\partial f(x)$, $\det A$ is nonnegative [nonpositive].*

With this in mind, we shall prove the theorem below. Whyburn [25] calls f *quasi-interior* if y lies in the interior of $f(U)$ for every y and every open U that contains a compact component of $f^{-1}(y)$. To see related results for C^1 maps, consult Church [4], [5], Plastock [19] and Sternberg and Swan [24].

THEOREM C. *A proper Lipschitzian map $f: R^n \to R^n$ having the constant sign property is onto and quasi-interior if the generalized derivative is invertible off a set S of measure (μ) zero.*

3. Lipschitz degree. The proof of Theorem C (below in §4) will call on properties of Brouwer degree for continuous maps and a degree function for Lipschitzian maps introduced in Pourciau [22]. If V is a bounded open subset of R^n, if $f: \overline{V} \to R^n$ is continuous and if y is not in $f(\text{bdy } V)$, then the Brouwer degree $\deg_B(f, V, y)$ is defined. For the definition and properties of the Brouwer degree, read Lloyd [17]. When f is C^1 and $f'(x)$ is invertible for each x in $P = \{x \in V: f(x) = y\}$, then P is discrete, by the classical (C^1) inverse function theorem, and hence finite, since $\overline{V}$ is compact. In this case there is a pleasant formula for the Brouwer degree:

$$\deg_B(f, V, y) = \begin{cases} \sum_{x \in P} & \operatorname{sign} \det f'(x), \\ 0 & \text{if } P \text{ is empty.} \end{cases}$$

This formula motivates the definition of a degree function for Lipschitzian f. When f is Lipschitzian and $\partial f(x)$ is invertible at each x in $P = \{x \in V: f(x) = y\}$, then P is discrete, by the inverse map theorem (§1), and hence finite. For any x in P, the sign

of $\det A$ is then constant as A runs through $\partial f(x)$. In this case we define the *Lipschitz degree* of f on V with respect to y:

$$\deg_L(f,V,y)=\begin{cases}\sum_{x\in P} \operatorname{sign}\det \partial f(x), \\ 0 \quad \text{if } P \text{ is empty.}\end{cases}$$

In Pourciau [22] it is shown that $\deg_B=\deg_L$ whenever the Lipschitz degree is defined.

4. Proof of Theorem C. We begin with an observation: if for some $\bar{y}$ we find that $\deg_B(f,V,\bar{y})\neq 0$ whenever V is an open ball containing $f^{-1}(\bar{y})$ and centered at the origin, then f is onto. To see this, choose any y in R^n and let γ stand for the line segment connecting $\bar{y}$ to y. As f is proper and γ is compact, $f^{-1}(\gamma)$ is compact. We can therefore find a ball $V=\{x:|x|<\delta\}$ containing $f^{-1}(\gamma)$. Now by the homotopy invariance of degree, if $\gamma(t)=(1-t)\bar{y}+ty$ for $0\leq t\leq 1$, then $\deg_B(f,V,y)=\deg_B(f,V,\gamma(t))=\deg_B(f,V,\bar{y})\neq 0$. By properties of degree this implies $f(x)=y$ for some x in V, so f is onto.

Thus to prove the first assertion of Theorem C, that f is onto, we must produce such a point $\bar{y}$. Since S has measure (μ) zero, the generalized derivative $\partial f(x)$ must be invertible at some $\bar{x}$. By the inverse map theorem, the image $f(R^n)$ must then contain an open set W. Now $f(S)$ has measure(μ) zero, for S has measure(μ) zero and f is Lipschitzian, so there is some $\bar{y}$ in W that is not in $f(S)$. Then $f^{-1}(\bar{y})$ is nonempty, and it is compact because f is proper. Suppose V is any open ball centered at the origin and containing $f^{-1}(\bar{y})$. As $\bar{y}$ is not in $f(\operatorname{bdy} V)$, both $\deg_B(f,V,\bar{y})$ and $\deg_L(f,V,\bar{y})$ are defined. We have

$$\deg_B(f,V,\bar{y})=\deg_L(f,V,\bar{y})=\sum_{x\in f^{-1}(\bar{y})} \operatorname{sign}\det \partial f(x).$$

Because $\partial f(x)$ is invertible for each x in $f^{-1}(\bar{y})$, $\operatorname{sign}\det \partial f(x)\neq 0$ for these x. But then the constant sign property implies the sum

$$\sum_{x\in f^{-1}(\bar{y})} \operatorname{sign}\det \partial f(x)$$

is nonzero. Thus $\deg_B(f,V,\bar{y})\neq 0$, and this completes the proof that f is onto.

Now we show that f is quasi-interior. Choose any y in the image $f(R^n)$ and suppose C is a component of $f^{-1}(y)$. Because f is proper, C is compact. Let U be an open set containing C. We can then find a bounded open set V containing C with $\bar{V}\subset U$ and $f^{-1}(y)\cap \operatorname{bdy} V=\varnothing$. Now the Brouwer degree $\deg_B(f,V,y)$ is defined. By the properties of degree and the properness of f, there is an open neighborhood W of y, disjoint from $f(\operatorname{bdy} V)$, on which $\deg_B(f,V,\cdot)$ is constant. Since $f(S)$ has measure(μ) zero, we can choose some $\bar{w}$ which is in W but not $f(S)$. Then $\deg_L(f,V,\bar{w})$ is defined and

$$\deg_B(f,V,\bar{w})=\deg_L(f,V,\bar{w})=\sum_{x\in P} \operatorname{sign}\det \partial f(x),$$

where $P=\{x\in V: f(x)=\bar{w}\}$. Notice P is nonempty because f is onto. As the generalized derivative $\partial f(x)$ is invertible for each x in P, $\operatorname{sign}\det \partial f(x)$ is nonzero for these x. But now the constant sign property implies the sum

$$\sum_{x\in P} \operatorname{sign}\det \partial f(x)$$

is nonzero. Therefore $\deg_B(f, V, \bar{w}) \neq 0$. Yet $\deg_B(f, V, \cdot)$ is constant on W, so $\deg_B(f, V, w) \neq 0$ for all w in W, and this implies $W \subset f(V)$. Thus y lies in the interior of $f(U)$, which tells us f is quasi-interior.

5. Remark. The proof of Theorem C presented here depends on the interior of $f(S)$ being empty. The assumption $\mu(S)=0$ ensures that the interior of $f(S)$ is empty, but this assumption may be unnecessary. Said differently, if $f: R^n \to R^n$ is Lipschitzian and $S=\{x: \partial f(x)$ contains a singular linear map$\}$, must the interior of $f(S)$ be empty? This is a Sard-like question. The answer to the C^1-analogue is trivially yes: if f is C^1 and $T=\{x: f'(x)$ is singular$\}$, then the interior of $f(T)$ is empty, indeed $f(T)$ has measure (μ) zero, because

$$\mu[f(T)] = \int_T \det f'(x)\, d\mu(x).$$

When f is Lipschitzian, this argument still works to tell us $f(T)$ has measure(μ) zero, yet it tells us nothing about $f(S)$. Of course, $S=T$ when f is C^1, but not when f is Lipschitzian.

REFERENCES

[1] K. J. Arrow and G. Debreu, *Existence of an equilibrium for a competitive economy*, Econometrica, 22 (1954), pp. 265–290.

[2] M. S. Berger and N. G. Meyers, *Generalized differentiation and utility functions*, in Preference, Utility and Demand, Harcourt, New York, 1971.

[3] L. O. Chua and Y. F. Lam, *Global homeomorphisms of vector-valued functions*, J. Math. Anal. Appl., 39 (1972), pp. 600–624.

[4] P. T. Church, *Differentiable open maps on manifolds*, Trans. Amer. Math. Soc., 109 (1963), pp. 87–100.

[5] _____, *Differentiable maps with non-negative Jacobian*, J. Math. Mech., 16 (1967), pp. 703–708.

[6] P. T. Church and E. Hemmingsen, *Light open maps on n-manifolds*, Duke Math. J., 27 (1960), pp. 527–536.

[7] F. H. Clarke, *Necessary conditions for nonsmooth problems in optimal control and calculus of variations*, Ph.D. dissertation, Univ. Washington, Pullman, 1973.

[8] _____, *On the inverse function theorem*, Pacific J. Math., 64 (1976), pp. 97–102.

[9] E. Dierker, *Two remarks on the number of equilibria of an economy*, 40 (1972), pp. 951–953.

[10] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[11] T. Fujisawa and E. S. Kuh, *Piecewise linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.

[12] J. Hadamard, *Sur les transformatinos ponctuelles*, Bull. Soc. Math. France, 34 (1906), pp. 71–84.

[13] H. Haneda, *Generalization of monotonicity applied to the DC analysis of nondifferentiable networks*, IEEE Trans. Circuits and Systems, CAS-Z1 (1974), pp. 406–412.

[14] J-B. Hiriart-Urruty, *On necessary optimality conditions in nondifferentiable programming*, Math. Programming, 14 (1978), pp. 73–86.

[15] Y. Kawamura, *Invertibility of Lipschitz continuous mappings and its applications to electrical network equations*, this Journal, 10 (1979), pp. 253–265.

[16] M. Kojima and R. Saigal, *A study of PC^1 homeomorphisms on subdivided polyhedrons*, this Journal, 10 (1979), pp. 1299–1312.

[17] N. G. Lloyd, *Degree Theory*, Cambridge Univ. Press, Cambridge, 1978.

[18] R. S. Palais, *Natural operations on differential forms*, Trans. Amer. Math. Soc., 92 (1959), pp. 125–141.

[19] R. A. Plastock, *Nonlinear Fredholm maps of index zero and their singularities*, Proc. Amer. Math. Soc., 68 (1978), pp. 317–322.

[20] B. H. Pourciau, *Analysis and optimization of Lipschitz continuous mappings*, J. Optim. Theory Appl., 22 (1977), pp. 311–351.

[21] _____, *Hadamard's theorem for locally Lipschitz maps*, J. Math. Anal. Appl., 84 (1981), pp. 279–285.

[22] _____, *Univalence and degree for Lipschitz continuous maps*, Arch. Rational Mech. Anal., to appear.

[23] _____, *Homeomorphisms and generalized derivatives*, J. Math. Anal. Appl., to appear.

[24] S. Sternberg and R. G. Swan, *On maps with nonnegative Jacobian*, Michigan Math. J., 6 (1959), pp. 339–342.

[25] G. T. Whyburn, *Analytic Topology*, AMS Colloquium Publications, vol. 28, American Mathematical Society, New York, 1942.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July, 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0013 $01.25/0

MEAN CONVERGENCE AND INTERPOLATION IN ROOTS OF UNITY*

A. SHARMA† AND P. VERTESI‡

Dedicated to Professor I. J. Schoenberg on his 78th birthday

Abstract. In 1940, Lozinski [Mat. Sb. (N.S.), 8(50) (1940), pp. 57–68] showed that if $f(z)$ is analytic in $|z|<1$ and continuous in $|z|\leq 1$, then the Lagrange interpolants to f in the nth roots of unity converge to $f(z)$ in the p-norm on the circle as $n\to\infty$, $p>0$. Here we provide a different proof of this result and extend a recent result of Saff and Walsh [Pacific J. Math., 45 (1973), pp. 639–641] in the same spirit.

Key words. Lagrange interpolation, mean convergence, roots of unity

1. Introduction. Let $S_n=\{z_{1n},z_{2n},\cdots,z_{n,n}\}$ denote a set of n points for each positive integer n and let $L_n(f;z)$ denote the Lagrange interpolation polynomial for a given function $f(z)$ defined on S_n. Then

$$L_n(f;z)=\sum_{k=1}^{n} f(z_{k,n})l_{kn}(z), \qquad l_{kn}(z)=\frac{\omega(z)}{(z-z_{kn})\omega'(z_{kn})}, \tag{1}$$

where $\omega(z)=\Pi_1^n(z-z_{kn})$. In 1937, Erdös and Turán [4] showed that if $S_{2n+1}=\{2k\pi/(2n+1),k=0,1,\cdots,2n\}$ and if $I_{2n+1}(g;\theta)$ is the unique trigonometric polynomial which interpolates a real 2π-periodic continuous function $g(\theta)$ on S_{2n+1}, then $I_{2n+1}(g;\theta)$ converges to $g(\theta)$ in the sense of mean square convergence. Later Erdös and Feldheim [3] showed that even more is true if S_n is the set of Chebyshev abscissas and $f(x)$ is continuous on $[-1,1]$, viz.,

$$\int_{-1}^{1}|f(x)-L_n(f;x)|^p\frac{dx}{\sqrt{1-x^2}}=0, \qquad p>0. \tag{2}$$

In 1964, Walsh and Sharma [11] gave an analogue in the complex plane of the 1937 result of Erdös and Turán. An extension of this to analytic curves with an extensive bibliography is due to J. H. Curtiss [5]. More general results, however, had been given much earlier by S. Lozinski [6] and Y. Alper [1].

Recently Saff and Walsh [10] have extended the result of Walsh and Sharma [11] to functions which are meromorphic in $|z|<1$, and continuous in $|z|\leq 1$ with precisely ν poles in $|z|<1$. If $r_{n\nu}(z)$ is the rational function of type (n,ν) which interpolates $f(z)$ in the $n+\nu+1$ roots of unity, they show that

$$\lim_{n\to\infty}\int_C|f(z)-r_{n\nu}(z)|^2|dz|=0. \tag{3}$$

For a detailed bibliography and general results on a real interval we refer to two significant papers of R. Askey [2] and G. P. Nevai [9]. There has been a recent revival of interest in such problems (see Varma and Vèrtesi [12]).

*Received by the editors September 8, 1981, and in revised form December 9, 1982.

†University of Alberta, Edmonton, Alberta, Canada T6G 2H1.

‡University of Alberta, Edmonton, Alberta, Canada T6G 2H1. The research of this author was supported by the National Research Council grant 3094.

The object of this note is to provide a different proof of the general result of Lozinski [6] by the method of Erdös and Feldheim [3]. We extend this result to linear operators which were studied by Motzkin and Sharma [8]. We also give a generalization of (3).

2. Statement of results. We shall prove

THEOREM 1. *Let $f(z)$ be analytic in $|z|<1$ and continuous in $|z|\leq 1$. If $L_n(f;z)$ denotes the Lagrange interpolation polynomial of degree $n-1$ which interpolates $f(z)$ in the n roots of unity, then*

$$\lim_{n\to\infty}\int_C |f(z)-L_n(f;z)|^p|dz|=0, \qquad p>0, \tag{4}$$

where C is the unit circle $|z|=1$.

COROLLARY. *Let $z_1,z_2,\cdots,z_\nu$ be ν (not necessarily distinct) points in $|z|<1$. If $f(z)$ satisfies the conditions of Theorem 1 and if $L_n^{(\nu)}(f;z)$ is the polynomial of degree $n-1$ which interpolates $f(z)$ in $(n-\nu)$ roots of unity and in the points $z_1,z_2,\cdots,z_\nu$, then*

$$\lim_{n\to\infty}\int_C |f(z)-L_n^{(\nu)}(f;z)|^p|dz|=0. \tag{5}$$

For a given integer r, consider the linear operator $L_{n,r}(f;z)$ given by

$$L_{n,r}(f;z)=\sum_{k=0}^{n-1} f(\omega^k)l_{kr}(z), \qquad \omega^n=1 \tag{6}$$

where

$$l_{kr}(z)=\frac{\omega^{(r+1)k}}{n}\,\frac{z^{n-r}-\omega^{k(n-r)}}{z-\omega^k}. \tag{7}$$

These operators were studied by Motzkin and Sharma [8] in connection with polynomials of best approximation on the roots of unity. For $r=0$, $l_{k0}(z)$ is the fundamental polynomial of Lagrange interpolation of degree $n-1$ given by (10).

We shall prove

THEOREM 2. *If $f(z)$ is analytic in $|z|<1$ and continuous in $|z|\leq 1$, then we have*

$$\lim_{n\to\infty}\int_C |L_{n,r}(f;z)-f(z)|^p|dz|=0, \qquad p>0. \tag{8}$$

THEOREM 3. *If $f(z)$ is meromorphic in $|z|<1$ and continuous in $|z|\leq 1$ with precisely ν poles in $|z|<1$, let $r_{n\nu}(f;z)$ denote the rational function of type (n,ν)*

$$r_{n,\nu}(f;z)=p_{n\nu}(z)/q_{n\nu}(z), \qquad q_{n\nu}(z) \text{ monic},$$

which interpolates $f(z)$ in $(n+\nu+1)$ roots of unity. Then

$$\lim_{n\to\infty}\int_C |f(z)-r_{n\nu}(f;z)|^p|dz|=0, \qquad p>0. \tag{9}$$

Remark. For $p=2$, (4) was proved in [11] but the general case had been proved earlier by Lozinski [6]. An independent proof of (4) for Jordan arcs with $p=2$ was given by Curtiss [5], but the general case was done earlier by Y. Alper [1]. Formula (9) was proved for $p=2$ by Saff and Walsh [10].

3. Some lemmas. The fundamental polynomials $l_k(z)$ of Lagrange interpolation on the n roots of unity are given by

$$l_k(z)=\frac{z^n-1}{z-\omega^k}\frac{\omega^k}{n}. \tag{10}$$

We shall prove

LEMMA 1. *If ν_1,μ_i $(i=1,2,\cdots,s)$ are $2s$ distinct integers $<n$, then we have*

$$\int_C l_{\nu_1}(z)\cdots l_{\nu_s}(z)\overline{l_{\mu_1}(z)}\cdots\overline{l_{\mu_s}(z)}|dz|=0. \tag{11}$$

From (10) and (11), we get

COROLLARY 2. *If $0\le\theta<2\pi$ and if*

$$t_k(\theta)=\frac{\sin^{(n\theta/2)}}{n\sin^{((\theta-\theta_k)/2)}},\qquad \theta_k=\frac{2k\pi}{n}\qquad (k=0,1,\cdots,n-1), \tag{12}$$

then we have

$$\int_0^{2\pi} t_{\nu_1}(\theta)\cdots t_{\nu_{2s}}(\theta)\,d\theta=0,\qquad \nu_1\ne\nu_2\ne\cdots\ne\nu_{2s}. \tag{13}$$

This corollary follows immediately from (11), on observing that for $z=e^{i\theta}$, and $\omega^\nu=e^{i\theta_\nu}$, we have

$$l_\nu(z)=e^{i(n-1)\theta/2}\frac{e^{i\theta_\nu/2}}{n}\frac{\sin^{(n\theta/2)}}{\sin^{((\theta-\theta_\nu)/2)}}.$$

Proof of Lemma 1. Since $\nu_i\ne\nu_j$, $i\ne j$, we have

$$\frac{1}{\prod_1^s(z-\omega^{\nu_k})}=\sum_{k=1}^{s}\frac{A_k}{z-\omega^{\nu_k}},\qquad A_k=\prod_{\substack{j=1\\ j\ne k}}^{s}(\omega^{\nu_k}-\omega^{\nu_j})^{-1}.$$

It follows that

$$\prod_{j=1}^{s} l_{\nu_j}(z)=\frac{\omega^V(z^n-1)^{s-1}}{n^{s-1}}\sum_{k=1}^{s}A_k\omega^{-\nu_k}l_{\nu_k}(z),\qquad V=\sum_1^s\nu_k. \tag{14}$$

Similarly, we have

$$\prod_{j=1}^{s} l_{\mu_j}(z)=\frac{\omega^U(z^n-1)^{s-1}}{n^{s-1}}\sum_{k=1}^{s}B_k\omega^{-\mu_k}l_{\mu_k}(z),\qquad U=\sum_1^s\mu_k, \tag{15}$$

where

$$B_k=\prod_{\substack{j=1\\ j\ne k}}^{s}(\omega^{\mu_k}-\omega^{\mu_j})^{-1}.$$

It is known [11] and is also very easy to verify that

$$\int_C l_\nu(z)\overline{l_\mu(z)}|dz|=\begin{cases}0, & \mu\ne\nu,\\ \dfrac{2\pi}{n}, & \mu=\nu.\end{cases} \tag{16}$$

Since

$$(z^n-1)^{s-1}\overline{(z^n-1)}^{s-1}=\sum_{k=-(s-1)}^{s-1}c_k z^{kn},$$

where

$$c_0=\sum_{k=0}^{s-1}\left\{\binom{s-1}{k}\right\}^2=\binom{2s-2}{s-1},$$

and since

$$\int_C l_\nu(z)\overline{l_\mu(z)}z^{\lambda n}|dz|=0,\qquad \lambda\neq 0,$$

it follows from (16) that

$$\int_C l_\nu(z)\overline{l_\mu(z)}(z^n-1)^{s-1}\overline{(z^n-1)}^{(s-1)}|dz|=\begin{cases}0, & \mu\neq\nu,\\ \binom{2s-2}{s-1}\dfrac{2\pi}{n}, & \mu=\nu.\end{cases}\tag{17}$$

Formula (11) follows from (14), (15) and (17).

4. Proof of Theorem 1. In order to prove (4), it is enough to prove

$$\lim_{n\to\infty}\int_C|f(z)-L_n(f;z)|^{2r}d\theta=0,\tag{18}$$

where r is any positive integer. If $P_{n-1}(z)$ is the polynomial of best approximation to $f(z)$ on C, set

$$\Delta(z)=f(z)-P_{n-1}(z)\quad\text{and}\quad\max_{|z|=1}|\Delta(z)|=E_{n-1}(f).$$

Then from a known inequality [13, p. 93, formula (10)], we have (with $|dz|=d\theta$),

$$\begin{aligned}\int_C|f(z)-L_n(f;z)|^{2r}d\theta &\le 2^{2r-1}\int|\Delta(z)|^{2r}d\theta+2^{2r-1}\int|L_n(\Delta;z)|^{2r}d\theta\\ &\le 2^{2r-1}2\pi(E_{n-1}(f))^{2r}+2^{2r-1}\int_C|L_n(\Delta;z)|^{2r}d\theta.\end{aligned}$$

Moreover, we have

$$\begin{aligned}|L_n(\Delta;z)|^2&=\left|\sum_{k=0}^{n-1}\Delta(\omega^k)l_k(z)\right|^2\\ &=\left|\sum_{k=0}^{n-1}\Delta(\omega^k)e^{i\theta_k/2}t_k(\theta)\right|^2=S_{1,n}^2(\theta)+S_{2,n}^2(\theta),\end{aligned}$$

where

$$S_{1n}(\theta)=\sum_{k=0}^{n-1}A_k t_k(\theta),\qquad S_{2n}(\theta)=\sum_{k=0}^{n-1}B_k t_k(\theta)$$

and

$$A_k=\operatorname{Re}\Delta(\omega^k)e^{i\theta_k/2},\qquad B_k=\operatorname{Im}\Delta(\omega^k)e^{i\theta_k/2}.$$

For $k=0,1,\cdots,n-1$, we obviously have $|A_k|\le E_{n-1}(f)$, $|B_k|\le E_{n-1}(f)$. Hence we have (by [13, p. 93, formula (10)]),

$$\int_C|L_n(\Delta;z)|^{2r}d\theta\le 2^{r-1}\int_C|S_{1n}(\theta)|^{2r}d\theta+2^{r-1}\int_C|S_{2n}(\theta)|^{2r}d\theta.$$

Here the relations

$$\lim_{n\to\infty}\int_0^{2\pi}|S_{kn}(\theta)|^{2r}d\theta=0\qquad(k=1,2)$$

follow mutatis mutandis from the reasoning of Erdös and Feldheim [3] on using Corollary 2 and the fact that

$$\sum_{k=0}^{n-1}|t_k(\theta)|^{1+\varepsilon}=O(1),\qquad\varepsilon>0,\tag{19}$$

which can be verified from (12). This completes the proof of Theorem 1.

Proof of Corollary 1. If $p_\nu(z)$ is the polynomial which interpolates $f(z)$ at z_k, set

$$F(z)=[f(z)-p_\nu(z)]/\Pi(z),\qquad\Pi_\nu(z)=\prod_1^\nu(z-z_j).$$

Then $F(z)$ is analytic in $|z|<1$ and continuous in $|z|\le1$. If $P_{n-\nu}(z)$ is the polynomial which interpolates $F(z)$ in $(n-\nu)$ roots of unity, we have from Theorem 1

$$\lim_{n\to\infty}\int_C|F(z)-P_{n-\nu}(z)|^p d\theta=0,$$

which yields

$$\lim_{n\to\infty}\int_C\frac{|f(z)-p_\nu(z)-\Pi_\nu(z)p_{n-\nu}(z)|^p}{|\Pi_\nu(z)|^p}d\theta=0.$$

Since $L_n^{(\nu)}(f;z)=p_\nu(z)+\Pi_\nu(z)P_{n-\nu}(z)$ and since $\Pi_\nu(z)$ is bounded on C, we have (5).

5. Proof of Theorem 2. It follows from (7) that

$$l_{kr}(z)=l_{k0}(z)-\frac{\omega^k}{n}z^{n-1}-\frac{\omega^{2k}}{n}z^{n-2}-\cdots-\frac{\omega^{rk}}{n}z^{n-r},$$

where $l_{k0}(z)=l_k(z)$ is given by (10). Since

$$z^{n-\nu}=\sum_{p=0}^{n-1}\omega^{p(n-\nu)}l_{p0}(z),$$

we have

$$l_{kr}(z)=l_{k0}(z)-\frac{1}{n}\sum_{\nu=1}^{r}\omega^{\nu k}\sum_{p=0}^{n-1}\omega^{p(n-\nu)}l_{p0}(z)$$

$$=\sum_{p=0}^{n-1}\alpha_{pr}^{(k)}l_{p0}(z),$$

where

$$\alpha_{pr}^{(k)}=\begin{cases}1-\dfrac{r}{n} & \text{if } p=k,\\ -\dfrac{1}{n}\displaystyle\sum_{\nu=1}^{r}\omega^{\nu(k-p)} & \text{if } p\neq k.\end{cases} \tag{20}$$

From (6), we then have

$$L_{n,r}(f;z)=\sum_{k=0}^{n-1}f(\omega^k)l_{kr}(z)=\sum_{k=0}^{n-1}f(\omega^k)\sum_{p=0}^{n-1}\alpha_{pr}^{(k)}l_{p0}(z)=\sum_{p=0}^{n-1}\beta_{pr}l_{p0}(z), \tag{21}$$

where

$$\beta_{p,r}=\sum_{k=0}^{n-1}f(\omega^k)\alpha_{pr}^{(k)}. \tag{22}$$

It follows from (20) that $|\beta_{p,r}|\leq 2\max|f|$. Since $L_{n,r}(f;z)=f(z)$ if $f(z)$ is a polynomial of degree $\leq n-r-1$, we can follow the method of §4 to complete the proof of Theorem 2.

Remark. A slight variation on the above proof depends on the following estimate,

$$\int_C l_{\nu_1 r}(z)\cdots l_{\nu_s r}(z)\overline{l_{\mu_1 r}(z)\cdots l_{\mu_s r}(z)}|dz|=O(n^{-2s}). \tag{23}$$

The proof of (23) can be given on the lines of the proof of Lemma 1, but is more tedious.

Proof of Theorem 3. If $f(z)$ is meromorphic in $|z|<1$ and has precisely ν poles $\alpha_1,\alpha_2,\cdots,\alpha_\nu$ in $|z|<1$, then following Saff and Walsh [10], we set

$$Q_0(z)=1,\qquad Q_k(z)=\prod_{j=1}^{k}(z-\alpha_j),\qquad 1\leq k\leq\nu,$$

$$q_n(z)=Q_\nu(z)+\sum_{k=1}^{\nu}a_k^{(n)}Q_{k-1}(z).$$

In fact, $\lim_{n\to\infty}a_k^{(n)}=0$. For large n, Saff and Walsh have shown that there exists number $a_k^{(n)}$ such that $Q_\nu(z)$ divides $L_{n+\nu}(q_nQ_\nu f;z)$. Then

$$r_{n\nu}(z)\equiv L_{n+\nu}(q_nQ_\nu f;z)\Big/(q_n(z)Q_\nu(z))$$

is a rational function of type (n,ν) for large n. From Theorem 1, we have for $k=1,2,\cdots,\nu+1$

$$\lim_{n\to\infty}\int_C |L_{n+\nu}(Q_{k-1}Q_\nu f;z)-Q_{k-1}Q_\nu f(z)|^p d\theta=0, \qquad p>0,$$

so that since

$$\int_C |L_{n+\nu}(q_nQ_\nu f;z)-q_nQ_\nu f(z)|^p d\theta$$

$$\leq c_1\int_C |L_{n+\nu}(Q_\nu^2 f;z)-Q_\nu^2 f(z)|^p d\theta$$

$$+c_1\sum_{k=1}^{\nu}\int_C |L_{n+\nu}Q_{k-1}Q_\nu(f;z)-Q_{k-1}Q_\nu(f)|^p d\theta$$

the result follows immediately.

6. Final remarks. If we consider the trigonometric interpolatory polynomials $I_{2n+1}(g,\theta)$ based on the nodes $2k\pi(2n+1)^{-1}$, $k=0,1,\cdots,2n$, using Corollary 2, (16) and again the reasoning of [3], we can prove

THEOREM 3. *Let $g(\theta)$ be any 2π-periodic continuous function. Then*

$$\lim_{n\to\infty}\int_0^{2\pi} |I_{2n+1}(g,\theta)-g(\theta)|^p d\theta=0, \qquad p>0.$$

Theorem 3 was first proved by J. Marcinkiewicz [7]. Our proof is different from his.

REFERENCES

[1]. S. Y. ALPER AND G. I. KALINOGORSKAJA, *The convergence of Lagrange interpolation polynomials in the complex domain*. Izv. Akad. Nauk S.S.S.R. Ser. Mat., 19 (1955), pp. 423–444. (In Russian)

[2]. R. ASKEY, *Mean convergence of orthogonal series and Lagrange interpolation*, Acta Math. Sci. Hungar., 23 (1972), pp. 79–85.

[3]. P. ERDÖS, E. FELDHEIM, *Sur le mode de convergence dans l'interpolation de Lagrange*, C. R. Acad. Sci. Paris, 203 (1936), pp. 913–915.

[4]. P. ERDÖS AND P. TURÁN, *On interpolation* I,. Ann. Math., 38 (1937), pp. 142–155.

[5]. J. H. CURTISS, *Convergence of complex Lagrange interpolation polynomials on the locus of interpolation points*, Duke Math. J., 32 (1965), pp. 187–204.

[6]. S. M. LOZINSKI, *Über interpolation*, Mat. Sbornik (N.S.), 8(50) (1940), pp. 57–68. (In Russian)

[7]. J. MARCINKIEWICZ, *Sur l'interpolation* I, Studies Math., 6 (1936), pp. 1–17.

[8]. T. S. MOTZKIN AND A. SHARMA, *A sequence of linear polynomial operators and their approximation-theoretic properties*, J. Approx. Theory, 5 (1972), pp. 176–198.

[9]. G. P. NEVAI, *Mean convergence of Lagrange interpolation*, I, J. Approx. Theory, 18 (1976), pp. 363–377.

[10]. E. B. SAFF AND J. L. WALSH, *On the convergence of rational functions which interpolate in the roots of unity*, Pacific J. Math., 45 (1973), pp. 639–641.

[11]. J. L. WALSH AND A. SHARMA, *Least square approximation and interpolation in roots of unity*, Pacific J. Math., 14 (1964), pp. 727–730.

[12]. A. K. VARMA AND P. VÉRTESI, *Some Erdös-Feldheim type theorems on mean convergence of Lagrange interpolation*, Math. Anal. Appl., to appear.

[13]. J. L. WALSH, *Interpolation and Approximation*, AMS Colloquium Publications 20, American Mathematical Society, Providence, RI, 1960.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0014 $01.25

TURÁN INEQUALITIES FOR ULTRASPHERICAL AND CONTINUOUS q-ULTRASPHERICAL POLYNOMIALS*

JOAQUIN BUSTOZ† AND MOURAD E. H. ISMAIL†

Abstract. Paul Turán discovered that Legendre polynomials satisfy the inequality

$$P_n^2 - P_{n+1}P_{n-1} > 0 \quad \text{for } -1 < x < 1.$$

It was then discovered that such an inequality holds for Jacobi polynomials, generalized Laguerre polynomials, and Bessel functions. In this paper we prove that the continuous q-ultraspherical polynomials satisfy this inequality also.

Writing $F_n^\alpha(x) = P_n^\alpha(x)/P_n^\alpha(1)$ where $P_n^\alpha(x)$ is the nth ultraspherical polynomial we prove an inequality similar to the above,

$$(n+1)F_n^\alpha F_{n-1}^\beta - (n-1)F_{n+1}^\alpha F_{n-2}^\beta > 0, \qquad 0 < x < 1, \quad \tfrac{1}{2} < \alpha \le \beta \le \alpha + 1.$$

1. Introduction. It is well known that the classical orthogonal polynomials, as well as Bessel functions, satisfy inequalities of the form $Q_n^2 - Q_{n+1}Q_{n-1} > 0$ on the spectral interval. An inequality of this type is called a Turán inequality after Paul Turán who first observed it in Legendre polynomials. A somewhat similar type of inequality was studied by Bustoz and Savage in [3], [4]. These inequalities involve two parameters. For example, if $P_n^\alpha(x)$ is the ultraspherical polynomial defined by

$$(1 - 2xz + z^2)^{-\alpha} = \sum_{n=0}^{\infty} P_n^\alpha(x) z^n$$

and if $F_n^\alpha(x) = P_n^\alpha(x)/P_n^\alpha(1)$, then the following inequalities hold:

$$F_n^\alpha F_{n+1}^\beta - F_{n+1}^\alpha F_n^\beta > 0, \qquad -1 < x < 1, \quad -\tfrac{1}{2} < \alpha \le \beta \le \alpha + 1, \tag{1.1}$$

$$F_n^\alpha F_n^\beta - F_{n+1}^\alpha F_{n-1}^\beta > 0, \qquad -1 < x < 1, \quad -\tfrac{1}{2} < \alpha \le \beta \le \alpha + 1. \tag{1.2}$$

Note that (1.2) reduces to the usual Turán inequality when $\alpha = \beta$. In this paper we will prove another inequality along the lines of (1.1) and (1.2), namely,

$$(n+1)F_n^\alpha F_{n-1}^\beta - (n-1)F_{n+1}^\alpha F_{n-2}^\beta > 0, \qquad 0 < x < 1, \quad \tfrac{1}{2} < \alpha \le \beta \le \alpha + 1. \tag{1.3}$$

Notice that the indices in the second term of inequalities (1.1), (1.2) and (1.3) differ by one, two, and three respectively. It is worth noting that there can be no similar inequality with the indices differing by four or more since the roots of the polynomials will no longer interlace.

In §§3 and 4 we will prove a Turán inequality for the continuous q-ultraspherical polynomials. The proof runs along the lines of a proof by Otto Szász [9] for the case of ultraspherical polynomials. The continuous q-ultraspherical polynomials have a long history dating back to Rogers and Ramanujan and have been studied recently by Askey and Ismail [1], [2]. The polynomials are defined by the recursion

$$C_0(x;\beta|q) = 1, \qquad C_1(x;\beta|q) = 2x(1-\beta)(1-q)^{-1},$$

$$(1-q^{n+1})C_{n+1}(x;\beta|q) = 2x(1-\beta q^n)C_n(x;\beta|q) - (1-\beta^2 q^{n-1})C_{n-1}(x;\beta|q).$$

* Received by the editors October 20, 1981, and in revised form January 29, 1982. This research was partially supported by the National Science Foundation under grant MCS-8002539-1.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

The inequality to be proved is

$$\frac{(1-\beta)(1-q)}{(1-\beta^2 q)(1-\beta q^{n-1})}\left(\frac{q}{\beta}\right)^{n-1}\left(1-\beta^n E_n^2(x)\right)\le E_n^2(x)-E_{n+1}(x)E_{n-1}(x)$$

$$\le\frac{(1-q)(\beta q^2;q)_{n-1}(q;q)_{n-1}}{(q\beta^2;q)_n(\beta;q)_{n-1}},\qquad |x|\le\frac{1}{2}\left(\beta^{1/2}+\beta^{-1/2}\right),\quad 0<q<\beta<1$$

where $E_n(x)=(q;q)_n C_n(x;\beta|q)/(\beta^2;q)_n$. We shall always assume $0<q<1$.

2. Positivity of $S_n(x;\alpha,\beta)=(n+1)F_n^\alpha F_{n-1}^\beta-(n-1)F_{n+1}^\alpha F_{n-2}^\beta$. We will need the following identities. These identities and their proofs may be found in Rainville [8]. Usually we will suppress the independent variable and write F_n^λ for $F_n^\lambda(x)$.

$$(n+2\lambda)F_{n+1}^\lambda=2(n+\lambda)xF_n^\lambda-nF_{n-1}^\lambda,\qquad n\ge 1, \tag{2.1}$$

where $F_0^\lambda=1$, $F_1^\lambda=x$,

$$(1-x^2)\left(F_n^\lambda\right)'=n\left(F_{n-1}^\lambda-xF_n^\lambda\right),\qquad n\ge 1, \tag{2.2}$$

$$(n+2\lambda+1)(1-x^2)F_n^{\lambda+1}=(2\lambda+1)\left(F_n^\lambda-xF_{n+1}^\lambda\right),\qquad n\ge 0. \tag{2.3}$$

We will say that the roots of two polynomials $A(x)$ and $B(x)$ *interlace* if between every two roots of $A(x)$ there is precisely one root of $B(x)$ and vice versa.

LEMMA 2.1. *If $S_n(x;\alpha,\beta)>0$ for $0<x<1$ then the roots of F_n^α and F_{n-2}^β interlace.*

Proof. Let $x_1>x_2>\cdots$ be the positive roots of F_{n-2}^β. Then $S_n(x_j;\alpha,\beta)=(n+1)F_n^\alpha(x_j)F_{n-1}^\beta(x_j)$. Since $\operatorname{sgn}F_{n-1}^\beta(x_j)=(-1)^j$ and $S_n(x_j;\alpha,\beta)>0$ we have that $\operatorname{sgn}F_n^\alpha(x_j)=(-1)^j$. □

LEMMA 2.2. *Let*

$$C_n=\frac{(n-1)(2+\alpha-\beta)}{(n-1)(2+\alpha-\beta)+2\alpha+1}.$$

Then

$$\frac{d}{dx}\left[(1-x^2)^{\alpha+1/2}S_n(x;\alpha,\beta)\right] \tag{2.4}$$

$$=2\left[(\beta-\alpha-2)n+1-\alpha-\beta\right](1-x^2)^{\alpha-1/2}\left(F_{n-1}^\beta-C_nxF_{n-2}^\beta\right)F_{n+1}^\alpha.$$

Proof. Solve for nF_{n-1}^λ in (2.1) and replace in (2.2) to get

$$(1-x^2)\left(F_n^\lambda\right)'=(n+2\lambda)\left(xF_n^\lambda-F_{n+1}^\lambda\right). \tag{2.5}$$

Using (2.2) and (2.5) we find after simplifying $(1-x^2)S_n'=(2\alpha+1)xS_n+2[(\beta-\alpha-2)n+1-\alpha-\beta](F_{n-1}^\beta-C_nxF_{n-2}^\beta)F_{n+1}^\alpha$. This is equivalent to (2.4). □

LEMMA 2.3.

$$S_n(x;\alpha,\alpha+2)>0\quad \textit{for } 0<x<1,\quad \alpha>-\tfrac{1}{2},\ n=2,3,\cdots.$$

Proof. When $\beta=\alpha+2$ then $C_n=0$ and (2.4) becomes

$$\frac{d}{dx}\left[(1-x^2)^{\alpha+1/2}S_n(x;\alpha,\alpha+2)\right]=-2(2\alpha+1)(1-x^2)^{\alpha-1/2}F_{n-1}^{\alpha+2}F_{n+1}^\alpha. \tag{2.6}$$

According to (2.6) $S_n(x;\alpha,\alpha+2)$ is positive in $(0,1)$ if $S_n(x;\alpha,\alpha+2)$ is positive at the roots of $F_{n-1}^{\alpha+2}=0$ and $F_{n+1}^\alpha=0$. We consider two cases.

Case 1. Suppose $F_{n-1}^{\alpha+2}=0$. In this case $S_n=-(n-1)F_{n+1}^{\alpha}F_{n-2}^{\alpha+2}$. Replace λ by $\alpha+2$ and n by $n-1$ in (2.1) and use the fact that $F_{n-1}^{\alpha+2}=0$ to get

$$F_{n-2}^{\alpha+2}=-\frac{n+2\alpha+3}{n-1}F_n^{\alpha+2}.$$

Hence at points where $F_{n-1}^{\alpha+2}=0$ we have

$$S_n=(n+2\alpha+3)F_{n+1}^{\alpha}F_n^{\alpha+2}. \tag{2.7}$$

Applying the identity $(2\alpha+1)F_n^{\alpha}=(n+2\alpha+1)F_n^{\alpha+1}-nxF_{n-1}^{\alpha+1}$ twice we get

(2.8)

$$(2\alpha+1)(2\alpha+3)F_{n+1}^{\alpha}=(n+2\alpha+2)(n+2\alpha+4)F_{n+1}^{\alpha+2}-(n+1)(2n+4\alpha+5)xF_n^{\alpha+2}.$$

Replacing λ by $\alpha+2$ and n by $n+1$ in (2.1) and using the fact that $F_{n-1}^{\alpha+2}=0$ gives

$$F_{n+1}^{\alpha+2}=\frac{2(n+\alpha+2)x}{n+2\alpha+4}F_n^{\alpha+2}. \tag{2.9}$$

Replacing (2.9) in (2.8) finally gives

$$F_{n+1}^{\alpha}=(2\alpha+3)^{-1}(n+2\alpha+3)xF_n^{\alpha+2}. \tag{2.10}$$

Then replacing (2.10) in (2.7) shows that $S_n(x;\alpha,\alpha+2)>0$ at the roots of $F_{n-1}^{\alpha+2}=0$.

Case 2. Suppose $F_{n+1}^{\alpha}=0$. In this case $S_n=(n+1)F_n^{\alpha}F_{n-1}^{\alpha+2}$. From (2.3) we have

$$\begin{aligned}(n+2\alpha)(n+2\alpha-2)(n+2\alpha-1)(1-x^2)^2F_{n-1}^{\alpha+2}\\ =(2\alpha+3)(2\alpha+1)\left[(n+2\alpha-1)F_{n-1}^{\alpha}-(2n+4\alpha-3)xF_n^{\alpha}\right].\end{aligned} \tag{2.11}$$

Since $F_{n+1}^{\alpha}=0$ it follows from (2.1) that $F_{n-1}^{\alpha}=(2(n+\alpha)x/n)F_n^{\alpha}$, and hence (2.11) becomes

$$\begin{aligned}n(n+2\alpha)(n+2\alpha-2)(n+2\alpha-1)(1-x^2)^2F_{n-1}^{\alpha+2}\\ =(2\alpha+3)(2\alpha+1)\left[(2\alpha+1)n+(2\alpha-1)\alpha\right]xF_n^{\alpha}.\end{aligned} \tag{2.12}$$

Hence $S_n=(n+1)F_n^{\alpha}F_{n-1}^{\alpha+2}$ is of the form $S_n=A_nx(F_{n-1}^{\alpha+2})^2$ with $A_n>0$. This completes the proof of the lemma. □

LEMMA 2.4.

$$S_n(x;\alpha,\alpha+1)>0 \quad \textit{for } 0<x<1, \quad \alpha>-\tfrac{1}{2}, \quad n=2,3,\cdots.$$

Proof. When $\beta=\alpha+1$ then (2.4) becomes

$$\begin{aligned}\frac{d}{dx}\left[(1-x^2)^{\alpha+1/2}S_n(x;\alpha,\alpha+1)\right]\\ =-2(n+2\alpha)(1-x^2)^{\alpha-1/2}\left[F_{n-1}^{\alpha+1}-\frac{n-1}{n+2\alpha}F_{n-2}^{\alpha+1}\right]F_{n+1}^{\alpha}.\end{aligned}$$

We distinguish two cases.

Case 1. Suppose $F_{n-1}^{\alpha+1}=\frac{n-1}{n+2\alpha}F_{n-2}^{\alpha+1}$. From the identity $(2\alpha+1)F_{n-1}^{\alpha}=(n+2\alpha)F_{n-1}^{\alpha+1}-(n-1)xF_{n-2}^{\alpha+1}$, we conclude that $F_{n-1}^{\alpha}=0$. In this case S_n becomes $S_n(x;\alpha,\alpha+1)=\frac{n-1}{n+2\alpha}[(n+1)xF_n^{\alpha}-(n+2\alpha)F_{n+1}^{\alpha}]F_{n-2}^{\alpha+1}$. Since $F_{n-1}^{\alpha}=0$ we have from (2.1) that $(n+2\alpha)F_{n+1}^{\alpha}=2(n+\alpha)xF_n^{\alpha}$ so that $S_n(x;\alpha,\alpha+1)=(-(n-1)(n+2\alpha-1)/(n+2\alpha))F_n^{\alpha}F_{n-2}^{\alpha+1}$. Now from the identity $(2\alpha+1)F_n^{\alpha}=(n+2\alpha)xF_{n-1}^{\alpha+1}-(n-1)F_{n-2}^{\alpha+1}$ we have that $(2\alpha+1)F_n^{\alpha}=(n-1)(x^2-1)F_{n-2}^{\alpha+1}$. Hence in this case $S_n(x;\alpha,\alpha+1)=((n-1)^2(n+2\alpha-1)/(n+2\alpha))(1-x^2)(F_{n-2}^{\alpha+1})^2>0$.

Case 2. Suppose $F_{n+1}^{\alpha}=0$. Then $S_n(x;\alpha,\alpha+1)=(n+1)F_n^{\alpha}F_{n-1}^{\alpha+1}$. Since $F_{n+1}^{\alpha}=0$ it follows from identity (2.1) that $2(n+\alpha)xF_n^{\alpha}=nF_{n-1}^{\alpha}$. Using this equality in the identity $(n+2\alpha)(1-x^2)F_{n-2}^{\alpha+1}=(2\alpha+1)(F_{n-1}^{\alpha}-xF_n^{\alpha})$ we conclude that $F_{n-1}^{\alpha+1}=((2\alpha+1)(1-x^2)^{-1}/n)xF_n^{\alpha}$ so that $S_n(x;\alpha,\alpha+1)=((n+1)(2\alpha+1)(1-x^2)^{-1}/n)x(F_n^{\alpha})^2$. This completes the proof of the lemma. □

LEMMA 2.5. *Let* $\Delta_n=F_n^{\alpha}F_{n-1}^{\alpha}-F_{n+1}^{\alpha}F_{n-2}^{\alpha}$. *Then* $\Delta_n>0$ *for* $0<x<1$, $\alpha>-\frac{1}{2}$, $n=2,3,\cdots$.

Proof. By using the identities $nF_{n-1}^{\alpha}=(1-x^2)(F_n^{\alpha})'+nxF_n^{\alpha}$, $(n+2\alpha)F_{n+1}^{\alpha}=(n+2\alpha)xF_n^{\alpha}-(1-x^2)(F_n^{\alpha})'$ and $n(n-1)F_{n-2}^{\alpha}=[2(n-1+\alpha)x^2-(n-1+2\alpha)]nF_n^{\alpha}+2(n-1+\alpha)(1-x^2)(F_n^{\alpha})'$, we can write

$$n(n-1)(n+2\alpha)\Delta_n=A_n(F_n^{\alpha})^2+B_nF_n^{\alpha}(F_n^{\alpha})'+C_n[(F_n^{\alpha})']^2$$

where $A_n=2(n+2\alpha)(n-1+\alpha)nx(1-x^2)$, $B_n=-2\alpha(1-x^2)[1+2(n-1+\alpha)x^2]$, $C_n=2(n-1+\alpha)x(1-x^2)^2$. Next, using the identity (which follows from the differential equation satisfied by the ultraspherical polynomials)

$$(1-x^2)^2[(F_n^{\alpha})']^2=(2\alpha+1)xF_n^{\alpha}(F_n^{\alpha})'-n(n+2\alpha)(F_n^{\alpha})^2-(1-x^2)(F_n^{\alpha})^2\left[\frac{(F_n^{\alpha})'}{(F_n^{\alpha})}\right]',$$

we get

$$\frac{1}{2}n(n-1)(n+2\alpha)(1-x^2)^{-1}(F_n^{\alpha})^{-2}\Delta_n \tag{2.13}$$

$$=[(n-1+\alpha a)x^2-\alpha]\frac{(F_n^{\alpha})'}{F_n^{\alpha}}-(n+\alpha-1)x(1-x^2)\left[\frac{(F_n^{\alpha})'}{F_n^{\alpha}}\right]'.$$

Letting $x_1>x_2>\cdots>x_n$ denote the roots of F_n^{α} and recalling the symmetry of these roots, we find

$$\frac{(F_n^{\alpha})'}{F_n^{\alpha}}=\sum_{k=1}^{[n/2]}\frac{2x}{x^2-x_k^2},\quad n \text{ even},\qquad \frac{(F_n^{\alpha})'}{F_n^{\alpha}}=\frac{1}{x}+\sum_{k=1}^{[n/2]}\frac{2x}{x^2-x_k^2},\quad n \text{ odd},$$

and differentiating gives

$$\left[\frac{(F_n^{\alpha})'}{F_n^{\alpha}}\right]'=-2\sum_{k=1}^{[n/2]}\frac{x^2+x_k^2}{(x^2-x_k^2)^2},\quad n \text{ even},$$

$$\left[\frac{(F_n^{\alpha})'}{F_n^{\alpha}}\right]'=-\frac{1}{x^2}-2\sum_{k=1}^{[n/2]}\frac{x^2+x_k^2}{(x^2-x_k^2)^2},\quad n \text{ odd}.$$

Replacing these sums in (2.13), we get for n even,

$$\frac{1}{2}n(n-1)(n+2\alpha)(1-x^2)^{-1}(F_n^{\alpha})^{-2}\Delta_n=2x\sum_{k=1}^{[n/2]}\frac{f_k(x)}{(x^2-x_k^2)^2}$$

and for n odd,

$$\frac{1}{2}n(n-1)(n+2\alpha)(1-x^2)^{-1}(F_n^{\alpha})^{-2}\Delta_n=\frac{n-1}{x}+2x\sum_{k=1}^{[n/2]}\frac{f_k(x)}{(x^2-x_k^2)^2},$$

where $f_k(x)=[n-1-2(n+\alpha-1)x_k^2]x^2+(n+2\alpha-1)x_k^2$. It is easy to see that $f_k(x)>0$ for $\alpha>-\frac{1}{2}$, $0<x<1$, $n=2,3,\cdots$. This proves the lemma. □

Note also that for odd n we have a positive lower bound for Δ_n, that is,

$$\Delta_n > \frac{2(1-x^2)(F_n^\alpha)^2}{n(n+2\alpha)x}, \qquad 0<x<1.$$

LEMMA 2.6.

$$S_n(x;\alpha,\alpha)>0 \quad \text{for } 0<x<1, \quad \alpha>-\tfrac{1}{2}, \quad n=2,3,\cdots.$$

Proof. Setting $\beta=\alpha$ in (2.4) gives

$$\frac{d}{dx}\left[(1-x^2)^{\alpha+1/2}S_n(x,\alpha,\alpha)\right] = -2(2n-1+\alpha+\beta)(1-x^2)^{\alpha-1/2}(F_{n-1}^\alpha - C_n x F_{n-2}^\alpha)F_{n+1}^\alpha \tag{2.14}$$

where $C_n=2(n-1)/(2n+2\alpha-1)$. We separate the critical points in (2.14) into two cases:

Case 1. Suppose $F_{n-1}^\alpha = C_n x F_{n-2}^\alpha$. Then

$$S_n=\left[(n+1)C_n x F_n^\alpha-(n-1)F_{n+1}^\alpha\right]F_{n-2}^\alpha. \tag{2.15}$$

From (2.1) we find that if $F_{n-1}^\alpha = C_n x F_{n-2}^\alpha$ then

$$F_n^\alpha=(n-1)\left[\frac{4(n+\alpha-1)x^2-(2n+2\alpha-1)}{(2n+2\alpha-1)(n+2\alpha-1)}\right]F_{n-2}^\alpha \quad \text{and}$$

$$F_{n+1}^\alpha=\frac{2(n-1)x\left[4(n+\alpha)(n+\alpha-1)x^2-(n+\alpha)(2n+2\alpha-1)-n(n+2\alpha-1)\right]F_{n-2}^\alpha}{(n+2\alpha)(n+2\alpha-1)(2n+2\alpha-1)}.$$

Replacing in (2.15) we find that

$$\frac{(2n+2\alpha-1)^2(n+2\alpha-1)(n+2\alpha)S_n}{2(n-1)^2x}=\left(A_n x^2+B_n\right)\left(F_{n-2}^\alpha\right)^2$$

where $A_n=-4(n+\alpha-1)[n^2+2(\alpha-1)n+\alpha(2\alpha-3)]$ and $B_n=(2n+2\alpha-1)(n+\alpha)(2n+2\alpha-3)$. Hence $S_n>0$ at the critical points $F_{n-1}^\alpha = C_n x F_{n-2}^\alpha$ if $x^2>-B_n/A_n$. Since $0<-B_n/A_n<1$, this does not fill out the complete interval $(0,1)$.

Next we will establish that $S_n>0$ in an interval $(b_n,1)$ with $0<b_n<-B_n/A_n$ and thus prove that $S_n>0$ at all the critical points in question in $(0,1)$. Recall Δ_n as defined in Lemma 2.5. We have

$$S_n(x,\alpha,\alpha)=(n-1)\Delta_n+2F_n^\alpha F_{n-1}^\alpha.$$

By Lemma 2.5, $\Delta_n>0$. Now at the critical points $F_{n-1}^\alpha = C_n x F_{n-2}^\alpha$ we have by (2.1) that $(n+2\alpha-1)F_n^\alpha=[2(n+\alpha-1)C_n x^2-(n-1)]F_{n-2}^\alpha$, so that

$$S_n=(n-1)\Delta_n+2\frac{\left[2(n+\alpha-1)C_n x^2-(n-1)\right]C_n x\left(F_{n-2}^\alpha\right)^2}{n+2\alpha-1}.$$

Now $2(n+\alpha-1)C_n x^2-(n-1)>0$ if $(2n+2\alpha-1)/4(n+\alpha-1)<x^2\le 1$. Since $(2n+2\alpha-1)/4(n+\alpha-1)<-A_n/B_n$, this proves that $S_n>0$ at the critical points $F_{n-1}^\alpha = C_n x F_{n-2}^\alpha$ in $(0,1)$ for $\alpha>-\frac{1}{2}$, $n=2,3,\cdots$.

Case 2. Suppose $F_{n+1}^\alpha=0$. Then $S_n=(n+1)F_n^\alpha F_{n-1}^\alpha$. From (2.1) we find in this case that $nF_{n-1}^\alpha=2(n+\alpha)xF_n^\alpha$. Hence $S_n=(2(n+\alpha)x/n)(F_n^\alpha)^2>0$ for $0<x<1$, $\alpha>-\frac{1}{2}$, and $n=2,3,\cdots$. This completes the proof of the lemma. □

In proving $S_n(x;\alpha,\beta)$ positive for the continuous range $\alpha\le\beta\le\alpha+2$, $0<x<1$, we will need positivity of the expression $D_n(x;\alpha,\beta)=F_n^\alpha F_{n+1}^\beta-F_{n+1}^\alpha F_n^\beta$, $0<x<1$, for $-\frac{1}{2}<\alpha\le\beta\le\alpha+2$. As mentioned in the introduction, D_n was proved positive in [5] for the range $-\frac{1}{2}<\alpha\le\beta\le\alpha+1$. The same proof will carry over to $\alpha+1\le\beta\le\alpha+2$ if we prove $D_n>0$ where $\beta=\alpha+2$. We do this in the next lemma.

LEMMA 2.7. *Set* $D_n=F_n^\alpha F_{n+1}^{\alpha+2}-F_{n+1}^\alpha F_n^{\alpha+2}$. *Then* $D_n>0$ *for* $0<x<1$, $\alpha>-\frac{1}{2}$, $n=0,1,2,\cdots$.

Proof. From [5, eq. (2.1)] we have

$$\frac{d}{dx}\left[(1-x^2)^{\alpha-1/2}D_n\right]=4(1-x^2)^{\alpha-3/2}F_{n+1}^{\alpha+2}\left(F_{n+1}^{\alpha+2}-xF_n^{\alpha+2}\right).$$

We have two sets of critical points which we consider in two cases:

Case 1. Suppose $F_{n+1}^{\alpha+2}=xF_n^{\alpha+2}$. Then $D_n=F_n^{\alpha+2}(xF_n^\alpha-F_{n+1}^\alpha)$. From the identities $(1-x^2)(F_n^\alpha)'=(n+2\alpha)(xF_n^\alpha-F_{n+1}^\alpha)$ and $(2\alpha+1)(F_n^\alpha)'=(n+2\alpha)F_{n-1}^{\alpha+1}$, we find $(2\alpha+1)D_n=(1-x^2)F_n^{\alpha+2}F_{n-1}^{\alpha+1}$. Next, from the identity $(2\alpha+3)F_{n-1}^{\alpha+1}=(n+2\alpha+2)F_{n-1}^{\alpha+2}-(n-1)xF_{n-2}^{\alpha+2}$ we obtain $(2\alpha+1)(2\alpha+3)D_n=(1-x^2)F_n^{\alpha+2}[(n+2\alpha+2)F_{n-1}^{\alpha+2}-(n-1)xF_{n-2}^{\alpha+2}]$. Setting $\lambda=\alpha+2$ in (2.1) we find that if $F_{n+1}^{\alpha+2}=xF_n^{\alpha+2}$, then $F_{n-1}^{\alpha+2}=xF_n^{\alpha+2}$ and $(n-1)F_{n-2}^{\alpha+2}=[2(n+\alpha+1)x^2-(n+2\alpha+3)]F_n^{\alpha+2}$. Hence at the critical points in question, $(2\alpha+1)(2\alpha+3)D_n=x(1-x^2)[2(n+\alpha+1)(1-x^2)+2\alpha+3](F_n^{\alpha+2})^2$, and $D_n>0$.

Case 2. Suppose $F_{n+1}^\alpha=0$. Then $D_n=F_n^\alpha F_{n+1}^{\alpha+2}$. By iterating the identity $(n+2\alpha+2)(1-x^2)F_{n+1}^{\alpha+1}=(2\alpha+1)(F_{n+1}^\alpha-xF_{n+2}^\alpha)$ we find that if $F_{n+1}^\alpha=0$ then

$$(n+2\alpha+1)(n+2\alpha+2)(n+2\alpha+3)(n+2\alpha+4)(2\alpha+1)^{-1}(2\alpha+3)^{-1}(1-x^2)^2F_{n+1}^{\alpha+2}$$
$$=(n+1)x\left[2(n+\alpha+2)(1-x^2)+2\alpha+1\right]F_n^\alpha.$$

This expresses $F_{n+1}^{\alpha+2}$ as a positive multiple of F_n^α (positive when $\alpha>-\frac{1}{2}$, $0<x<1$, $n=0,1,\cdots$) and hence $D_n>0$. This completes the proof of the lemma. □

By Lemma 2.7 and [5, proof of Thm. 2.1] we have:

THEOREM 2.1. *Let* $D_n(x;\alpha,\beta)=F_n^\alpha F_{n+1}^\beta-F_{n+1}^\alpha F_n^\beta$. *Then* $D_n(x;\alpha,\beta)>0$ *for* $0<x<1$, $n=0,1,2,\cdots$, *if* $-\frac{1}{2}<\alpha\le\beta\le\alpha+2$.

We can now state and prove

THEOREM 2.2. (a) *If* $\frac{1}{2}<\alpha\le\beta\le\alpha+1$ *then* $S_n(x;\alpha,\beta)>0$ *for* $0<x<1$, $n=2,3,\cdots$.

(b) *If* $\frac{3}{2}<\alpha\le\beta\le\alpha+2$ *then* $S_n(x;\alpha,\beta)>0$ *for* $0<x<1$, $n=2,3,\cdots$.

Proof. First we prove part (a). By Lemmas 2.4 and 2.6 we have $S_n(x;\alpha,\alpha)>0$ and $S_n(x;\alpha,\alpha+1)>0$. Suppose $\alpha<\beta<\alpha+1$. Write $\psi_{n-1}^\beta=F_{n-1}^\beta-C_nxF_{n-2}^\beta$. By Lemma 2.2 $[(1-x^2)^{\alpha+1/2}S_n(x;\alpha,\beta)]'=2[(\beta-\alpha-2)n+1-\alpha-\beta](1-x^2)^{\alpha-1/2}\psi_n^\beta F_{n+1}^\alpha$. We have two cases.

Case 1. Suppose $F_{n+1}^\alpha=0$. Then $S_n=(n+1)F_n^\alpha F_{n-1}^\beta$. If $x_1>x_2>\cdots$ are the roots of F_{n+1}^α in $(0,1)$ then $\operatorname{sgn}F_n^\alpha(x_j)=(-1)^j$. By Lemmas 2.1, 2.4 and 2.6, the roots of F_{n+1}^α interlace the roots of F_{n-1}^α and $F_{n-1}^{\alpha+1}$. Since the roots of F_{n-1}^β are monotone decreasing functions of β, it follows that the roots of F_{n+1}^α and F_{n-1}^β interlace for $\alpha<\beta<\alpha+1$. Hence $\operatorname{sgn}F_{n-1}^\beta(x_j)=(-1)^j$ for $\alpha<\beta<\alpha+1$ and $S_n(x_j)>0$.

Case 2. Suppose $\psi_n^\beta=0$. Then $S_n=[(n+1)C_nxF_n^\alpha-(n-1)F_{n+1}^\alpha]F_{n-2}^\beta$. Let $z_1>z_2>\cdots$ be the roots of ψ_n^β in $(0,1)$. Then $\operatorname{sgn}F_{n-2}^\beta(z_j)=(-1)^{j+1}$, and to complete the proof we need $\operatorname{sgn}[(n+1)C_nz_jF_n^\alpha(z_j)-(n-1)F_{n+1}^\alpha(z_j)]=(-1)^{j+1}$. Set $g_n^\lambda(x,\alpha,\beta)=(n+1)C_nxF_n^\lambda-(n-1)F_{n+1}^\lambda$. In this notation we need $\operatorname{sgn}g_n^\alpha(z_j)=(-1)^{j+1}$. Note that $C_nz_j=F_{n-1}^\beta(z_j)/F_{n-2}^\beta(z_j)$ so that $g_n^\lambda(z_j)=S_n(z_j,\lambda,\beta)/F_{n-2}^\beta(z_j)$. Since $S_n(x,\lambda,\beta)>0$ in $(0,1)$ if $\lambda=\beta$, $\lambda=\beta-1$ (by Lemmas 2.4 and 2.6), we have $\operatorname{sgn}g_n^\lambda(z_j)=(-1)^{j+1}$ if $\lambda=\beta$ or $\lambda=\beta-1$, where we must require $\beta>\frac{1}{2}$ by Lemma 2.4. It follows that g_n^β and $g_n^{\beta-1}$

vanish exactly once between each two positive roots of ψ_{n-1}^{β}. Let $y_1>y_2>\cdots$ denote the positive roots of g_n^{β} and let $x_1>x_2>\cdots$ denote the positive roots of $g_n^{\beta-1}$. We will prove that if $\beta-1<\lambda<\beta$ then $\operatorname{sgn} g_n^{\lambda}(y_j)=(-1)^{j+1}$ and $\operatorname{sgn} g_n^{\lambda}(x_j)=(-1)^{j+2}$; from this we can conclude that g_n^{λ} vanishes exactly once between x_j and y_j. Since $g_n^{\beta}(y_j)=0$ we have that $(n+1)C_n y_j=(n-1)F_{n+1}^{\beta}(y_j)/F_n^{\beta}(y_j)$, so that

$$g_n^{\lambda}(y_j)=(n-1)D_n(y_j;\lambda,\beta)/F_n^{\beta}(y_j), \tag{2.16}$$

where D_n is defined in Theorem 2.1. Now by Theorem 2.1, $D_n(y_j;\lambda,\beta)>0$ if $-\frac{1}{2}<\lambda<\beta<\lambda+1$. Hence we require $\beta>\frac{1}{2}$ so that $\beta-1<\lambda<\beta$ is consistent with the hypothesis of Theorem 2.1. From (2.16) we then conclude that $\operatorname{sgn} g_n^{\lambda}(y_j)=\operatorname{sgn} F_n^{\beta}(y_j)=(-1)^{j+1}$. Similarly since $g_n^{\beta-1}(x_j)=0$ we get

$$g_n^{\lambda}(x_j)=-(n-1)D_n(x_j;\beta-1,\lambda)/F_n^{\beta-1}(x_j). \tag{2.17}$$

From (2.17) and Theorem 2.1 we get $\operatorname{sgn} g_n^{\lambda}(x_j)=-\operatorname{sgn} F_n^{\beta-1}(x_j)=(-1)^{j+2}$. We have now proved that if $\beta>\frac{1}{2}$ and $\beta-1<\lambda<\beta$ then g_n^{λ} vanishes exactly once between each two roots of ψ_{n-1}^{β}, or that is, $\operatorname{sgn} g_n^{\lambda}(z_j)=(-1)^{j+1}$. This completes the proof of (a).

The proof of (b) is virtually identical so we omit it. We point out only that in proving (b), equation (2.17) involves $D_n(x,\beta-2,\lambda)$, which is positive by Theorem 2.1 if $-\frac{1}{2}<\beta-2<\lambda<\beta-1$; hence we need $\beta>\frac{3}{2}$ instead of $\beta>\frac{1}{2}$ as in (a). □

The inequality $\Delta_n(x;\alpha,\beta)=F_n^{\alpha}F_{n-1}^{\beta}-F_{n+1}^{\alpha}F_{n-2}^{\beta}>0$, $0<x<1$, is cleaner and more natural than the inequality stated as Theorem 2.2. Lemma 2.5 proves this inequality when $\beta=\alpha>-\frac{1}{2}$. The special case $\beta=\alpha=\frac{1}{2}$ was proved by Forsyth [6]. It is probably true that $\Delta_n(x;\alpha,\beta)>0$, $0<x<1$, $-\frac{1}{2}<\alpha\le\beta\le\alpha+1$, although we can only prove it at the end points $\alpha=\beta$ and $\beta=\alpha+1$. We state these cases in the next theorem.

THEOREM 2.3. *Let* $\Delta_n(x;\alpha,\beta)=F_n^{\alpha}F_{n-1}^{\beta}-F_{n+1}^{\alpha}F_{n-2}^{\beta}$. *Then* $\Delta_n(x;\alpha,\alpha)>0$ *and* $\Delta_n(x;\alpha,\alpha+1)>0$ *for* $0<x<1$, $n=2,3,\cdots$, *and* $\alpha>-\frac{1}{2}$.

Proof. $\Delta_n(x;\alpha,\alpha)>0$ was proved in Lemma 2.5. By using the identity $(n+2\lambda+1)(1-x^2)F_n^{\lambda+1}=(2\lambda+1)(F_n^{\lambda}-xF_{n+1}^{\lambda})$ and proceeding as in the proof of Lemma 2.5, we can write

$$n(n-1)(n+2\alpha)\Delta_n(x;\alpha,\alpha+1)=-(2\alpha+1)(1-x^2)(F_n^{\alpha})^2\left\{\frac{(F_n^{\alpha})'}{F_n^{\alpha}}+x\left[\frac{(F_n^{\alpha})'}{F_n^{\alpha}}\right]\right\}'. \tag{2.18}$$

Letting x_k denote the roots of F_n^{α}, we have from (2.18)

$$n(n-1)(n+2\alpha)\Delta_n(x;\alpha,\alpha+1)=4(2\alpha+1)x(1-x^2)(F_n^{\alpha})^2\sum_{k=1}^{[n/2]}\frac{x_k^2}{(x^2-x_k^2)^2}>0.$$

This completes the proof. □

3. The continuous q-ultraspherical polynomials. Recall that the continuous q-ultraspherical polynomials $C_n(x;\beta|q)$ satisfy the recurrence relation (Askey and Ismail [1], [2]).

$$(1-q^{n+1})C_{n+1}(x;\beta|q)=2x(1-\beta q^n)C_n(x;\beta|q)-(1-\beta^2q^{n-1})C_{n-1}(x;\beta|q) \tag{3.1}$$

when $n>0$, and the initial conditions

$$C_0(x;\beta|q)=1, C_1(x;\beta|q)=2x(1-\beta)(1-q)^{-1}. \tag{3.2}$$

The generating function

$$\sum_0^\infty C_n(\cos\theta;\beta|q)t^n=\frac{(t\beta e^{i\theta};q)_\infty(t\beta e^{-i\theta};q)_\infty}{(te^{i\theta};q)_\infty(te^{-i\theta};q)_\infty} \tag{3.3}$$

and the q-binomial theorem lead to the following representation of $C_n(\cos\theta;\beta|q)$:

$$C_n(\cos\theta;\beta|q)=\sum_0^n \frac{(\beta;q)_k(\beta;q)_{n-k}}{(q;q)_k(q;q)_{n-k}}\cos[(n-2k)\theta]. \tag{3.4}$$

This implies the inequalities

$$|C_n(\cos\theta;\beta|q)|\le C_n(1;\beta|q),\qquad 0\le\theta\le\pi,\quad \beta<1, \tag{3.5}$$

$$|C_n(x;\beta|q)|\le C_n\left(\frac{\beta^{1/2}+\beta^{-1/2}}{2};\beta|q\right),\qquad 2|x|\le\beta^{1/2}+\beta^{-1/2},\quad 0<\beta\le 1. \tag{3.6}$$

The value of $C_n((\beta^{1/2}+\beta^{-1/2})/2;\beta|q)$ is $(\beta^2;q)_n\beta^{-n/2}/(q;q)_n$, but unfortunately we cannot evaluate $C_n(1;\beta|q)$ in general when $q\neq 1$. Luckily, however, we shall need only the asymptotic behavior of the bound $C_n(1;\beta|q)$ as $n\to\infty$, which can be obtained from applying the asymptotic method of Darboux to (3.3) with $\theta=0$. A comparison function (Olver [7, p. 309]) is

$$\frac{(\beta;q)_\infty^2}{(1-t)^2(q;q)_\infty^2}+\frac{a}{1-t}$$

where a is a suitably chosen constant. This implies the existence of a constant b such that

$$C_n(1;\beta|q)=\frac{(\beta;q)_\infty^2}{(q;q)_\infty^2}\left[n+1+a+\frac{b}{n}+o(n^{-2})\right]. \tag{3.7}$$

The existence of a and b will be used in the sequel but their exact values are not important. The exact values of a and b, if needed, can be computed by first using Cauchy's formulas to represent $C_n(1;\beta|q)$ by a contour integral, then applying Laplace's method for contour integrals (Olver [7]). We next derive a recurrence relation for the sequence of polynomials $D_n^\beta(x)$ defined via

(3.8)

$$D_n^\beta(x)=(1-\beta^2q^{n-1})(1-\beta q^n)E_n^2(x)-(1-\beta^2q^n)(1-\beta q^{n-1})E_{n-1}(x)E_{n+1}(x)$$

where

$$E_n(x)=(q;q)_nC_n(x;\beta|q)/(\beta^2;q)_n. \tag{3.9}$$

The E_n's depend on β and q as well as x. It is straightforward to obtain

$$(1-\beta^2q^n)E_{n+1}(x)=2x(1-\beta q^n)E_n(x)-(1-q^n)E_{n-1}(x), \tag{3.10}$$

from (3.9) and (3.1). The multiplication of (3.10) by E_n yields

$$\begin{aligned}2x(1-\beta q^n)E_n^2&=\{(1-\beta^2q^n)E_{n+1}(x)+(1-q^n)E_{n-1}(x)\}E_n(x)\\&=(1-\beta^2q^{n-1})^{-1}\{(1-\beta^2q^n)E_{n+1}(x)+(1-q^n)E_{n-1}(x)\}\\&\quad\cdot\{2x(1-\beta q^{n-1})E_{n-1}(x)-(1-q^{n-1})E_{n-2}(x)\};\end{aligned}$$

hence

$$
\begin{aligned}
&2x(1-\beta q^n)(1-\beta^2 q^{n-1})E_n^2(x)\\
&\quad=2x(1-\beta q^{n-1})(1-q^n)E_{n-1}^2(x)\\
&\qquad+2x(1-\beta q^{n-1})(1-\beta^2 q^n)E_{n-1}(x)E_{n+1}(x)\\
&\qquad-(1-q^{n-1})E_{n-2}(x)\{(1-\beta^2 q^n)E_{n+1}(x)+(1-q^n)E_{n-1}(x)\}\\
&\quad=2x(1-\beta q^{n-1})(1-q^n)E_{n-1}^2(x)+2x(1-\beta q^{n-1})(1-\beta^2 q^n)E_{n-1}(x)E_{n+1}(x)\\
&\qquad-(1-q^{n-1})E_{n-2}(x)2x(1-\beta q^n)E_n(x).
\end{aligned}
$$

Taking (3.8) into account we can rewrite the above identity as

$$
\begin{aligned}
D_n^\beta(x)&=(1-\beta q^{n-1})(1-q^n)E_{n-1}^2(x)-(1-q^{n-1})(1-\beta q^n)E_{n-2}(x)E_n(x)\\
&=(1-q^n)(1-\beta q^{n-1})E_{n-1}^2(x)+\frac{(1-q^{n-1})(1-\beta q^n)}{(1-\beta^2 q^{n-1})(1-\beta q^{n-2})}\\
&\qquad\cdot\{D_{n-1}^\beta(x)-(1-\beta^2 q^{n-2})(1-\beta q^{n-1})E_{n-1}^2(x)\}\\
&=\frac{(1-q^{n-1})(1-\beta q^n)}{(1-\beta^2 q^{n-1})(1-\beta q^{n-2})}D_{n-1}^\beta(x)+\frac{(1-\beta q^{n-1})p(\beta)}{(1-\beta^2 q^{n-1})(1-\beta q^{n-2})}E_{n-1}^2(x),
\end{aligned}
$$

where

$$p(\beta)=(1-\beta)(q-\beta)\left[q^{n-2}(1-q)+q^{2n-3}(1-q)\beta\right].$$

The above manipulations lead to

$$
\begin{aligned}
(3.11)\quad &\frac{(\beta^2;q)_n(\beta;q)_{n-1}}{(\beta q;q)_n(q;q)_{n-1}}D_n^\beta(x)\\
&=\frac{(\beta^2;q)_{n-1}(\beta;q)_{n-2}}{(\beta q;q)_{n-1}(q;q)_{n-2}}D_{n-1}^\beta(x)\\
&\qquad+\frac{q^{n-1}(1-\beta)^2(1-\beta/q)(1-q)(\beta^2;q)_{n-1}}{(1-\beta q^{n-2})(1-\beta q^n)}(1+\beta q^{n-1})E_{n-1}^2(x),
\end{aligned}
$$

holding for $n\geq 2$. Set

$$(3.12)\qquad \zeta_n^\beta(x)=\frac{(\beta^2;q)_n(\beta;q)_{n-1}}{(\beta q;q)_n(q;q)_{n-1}}D_n^\beta(x),$$

$$(3.13)\qquad g_n=q^n(1+\beta q^n)(\beta^2;q)_n\{(1-\beta q^{n+1})(1-\beta q^{n-1})\}^{-1},\qquad n>1.$$

Expressing the D's in (3.11) in terms of the ζ's we discover, upon iterating (3.11), the following relationship:

$$(3.14)\qquad \zeta_n^\beta(x)=\zeta_1^\beta(x)+(1-\beta)^2(1-\beta/q)(1-q)\sum_2^n g_{k-1}E_{k-1}^2(x).$$

The series $\sum_1^\infty g_k E_k^2(x)$ converges absolutely and uniformly on $[-1,1]$ because

$$\sup|E_k(x)|=O(C_n(1;\beta|q))=O(n)\quad\text{as } n\to\infty,\quad -1\leq x\leq 1$$

(see (3.7) and (3.9)) and $g_n = O(q^n)$ as $n \to \infty$ (see (3.13)). Furthermore, when $\beta > q$, $E_n((\beta^{1/2}+\beta^{-1/2})/2) = \beta^{-n/2}$ and the series $\sum_1^\infty g_k E_k^2(x)$ converges absolutely and uniformly on the interval $[-(\beta^{1/2}+\beta^{-1/2})/2, (\beta^{1/2}+\beta^{-1/2})/2]$.

4. The Turán inequality for the q-ultraspherical polynomials. The existence of $\lim_{n\to\infty} \zeta_n^\beta(x)$ follows from (3.14) and the convergence of the series on its right side. We now evaluate this limit explicitly and also compute $\zeta_1^\beta(x)$. It is obvious that

$$\lim_{n\to\infty} \zeta_n^\beta(x) = \frac{(\beta^2;q)_\infty(\beta;q)_\infty}{(\beta q;q)_\infty(q;q)_\infty} \lim_{n\to\infty} D_n^\beta(x). \tag{4.1}$$

When $x=1$, the limit $\lim_{n\to\infty} D_n^\beta(1)$ can be evaluated by using (3.7), (3.8) and (3.9). An easy calculation gives

$$\lim_{n\to\infty} D_n^\beta(1) = \left\{ \frac{(\beta;q)^2}{(q;q)_\infty(\beta^2;q)_\infty} \right\}^2;$$

hence

$$\lim_{n\to\infty} \zeta_n^\beta(1) = \frac{(1-\beta)(\beta;q)_\infty^4}{(\beta^2;q)_\infty(q;q)_\infty^3}. \tag{4.2}$$

The q-ultraspherical polynomials are symmetric polynomials in x, that is, $C_n(-x;\beta|q) = (-1)^n C_n(x;\beta|q)$. Hence $D_n^\beta(x)$ is an even function of x. It thus suffices to restrict x to the interval $[0,1]$. When $x \in (-1,1)$, the asymptotic behavior of $C_n(x;\beta\,|\,q)$ is (Askey and Ismail [2]),

$$C_n(\cos\theta;\beta|q) \sim 2\left| \frac{(\beta;q)_\infty(\beta e^{2i\theta};q)_\infty}{(q;q)_\infty(e^{2i\theta};q)_\infty} \right| \cos(n\theta-\phi), \qquad 0<\theta<\pi, \tag{4.3}$$

where

$$\phi = \arg\{(\beta;q)_\infty(\beta e^{2i\theta};q)_\infty/(e^{2i\theta};q)_\infty\}.$$

This establishes, via (3.8) and (3.9),

$$D_n^\beta(\cos\theta) \sim 4\left| \frac{(\beta;q)_\infty(\beta e^{2i\theta};q)_\infty}{(\beta^2;q)_\infty(e^{2i\theta};q)_\infty} \right|^2 \sin^2\theta, \qquad n\to\infty, \quad \theta\in(0,\pi), \tag{4.4}$$

which when combined with (4.1) shows that $\lim_{n\to\infty} \zeta_n^\beta(x)$ exists and

$$\lim_{n\to\infty} \zeta_n^\beta(\cos\theta) = \zeta^\beta(x) := \frac{(1-\beta)(\beta;q)_\infty^2(\beta e^{2i\theta};q)_\infty(\beta e^{-2i\theta};q)_\infty}{(q;q)_\infty(\beta^2;q)_\infty(qe^{2i\theta};q)_\infty(qe^{-2i\theta};q)}, \tag{4.5}$$

holds when $0<\theta<\pi$. When $\theta = 0, \pi$ formula (4.2) implies the validity of (4.5) at those two points.

We now compute $\zeta_1^\beta(x)$. The polynomial $C_2(x;\beta|q)$ is

$$C_2(x;\beta|q) = \frac{4(1-\beta)(1-\beta q)}{(1-q)(1-q^2)}x^2 - \frac{(1-\beta^2)}{(1-q^2)};$$

hence

$$\zeta_1^\beta(x) = (1-\beta)(1-q)(1-\beta^2)/(1-\beta q). \tag{4.6}$$

Combining (3.14), (4.5), and (4.6), we obtain

(4.7)

$$\zeta^{\beta}(x)=(1-\beta)(1-\beta^2)(1-q)/(1-\beta q)+(1-\beta)^2(1-\beta/q)(1-q)\sum_{1}^{\infty} g_k E_k^2(x).$$

We now consider the cases $\beta>q$ and $\beta<q$ separately.

Case 1. $\beta>q$. In this case the inequality

$$D_n^{\beta}(x)\geq D_n^{\beta}\left(\frac{\beta^{1/2}+\beta^{-1/2}}{2}\right), \qquad |x|\leq\frac{1}{2}(\beta^{1/2}+\beta^{-1/2}),$$

follows from (3.6), (3.14) and (3.12). The above inequality can be rewritten as

$$E_n^2(x)-E_{n+1}(x)E_{n-1}(x)\geq\frac{(1-\beta)(1-q)}{(1-\beta^2q^n)(1-\beta q^{n-1})}\left(\frac{q}{\beta}\right)^{n-1}(1-\beta^n E_n^2(x)), \tag{4.8}$$

$$|x|<\frac{1}{2}(\beta^{1/2}+\beta^{-1/2}), \qquad \beta>q.$$

Clearly the right member of (4.8) is nonnegative for $q<\beta<1$, $|x|<\frac{1}{2}(\beta^{1/2}+\beta^{-1/2})$. Another inequality follows from observing that $\zeta_n^{\beta}(x)$ is a decreasing sequence, so $\zeta_n^{\beta}(x)\leq\zeta_1^{\beta}(x)$. This establishes

$$E_n^2(x)-E_{n+1}(x)E_{n-1}(x)<\frac{(1-q)(1-\beta q^n)(q;q)_{n-1}}{(q\beta^2;q)_n(1-\beta q)}, \tag{4.9}$$

$$|x|\leq\frac{1}{2}(\beta^{1/2}+\beta^{-1/2}), \qquad \beta>q,\ n>1.$$

Both sides of (4.9) tend to a constant as $n\to\infty$ when $x=\pm1$ (see (3.7)). Although the constants are different, the right member of (4.9) is the correct order of magnitude. This establishes the following result:

THEOREM 4.1. *We have*

$$\frac{(1-q)(1-\beta q^n)(q;q)_{n-1}}{(q\beta^2;q)_n(1-\beta q)}\geq E_n^2(x)-E_{n+1}(x)E_{n-1}(x) \tag{4.10}$$

$$\geq\frac{(1-\beta)(1-q)}{(1-\beta^2q^n)(1-\beta q^{n-1})}\left(\frac{q}{\beta}\right)^{n-1}(1-\beta^n E_n^2(x)),$$

holding for $|x|\leq\frac{1}{2}(\beta^{1/2}+\beta^{-1/2})$, $\beta>q$.

Case 2. $\beta<q$. In this case $\zeta_n^{\beta}(x)$ increases as n increases. Thus $\zeta_n^{\beta}(x)\geq\zeta_1^{\beta}(x)$. Moreover, $\zeta_n^{\beta}(x)\leq\zeta_n^{\beta}((\beta^{1/2}+\beta^{-1/2})/2)$ for $|x|\leq\frac{1}{2}(\beta^{1/2}+\beta^{-1/2})$. These inequalities prove:

THEOREM 4.2. *The inequalities*

$$E_n^2(x)-E_{n+1}(x)E_{n-1}(x) \tag{4.11}$$

$$\geq\frac{(1-q)(1-\beta q^n)(q;q)_{n-1}}{(1-\beta q)(q\beta^2;q)_n}-\frac{(1-q)(1-\beta)\beta q^{n-1}}{(1-\beta q^{n-1})(1-\beta^2q^n)}E_n^2(x)$$

and

$$E_n^2(x)-E_{n+1}(x)E_{n-1}(x)\leq\frac{(1-\beta)(1-q)}{(1-\beta^2q^n)(1-\beta q^{n-1})}\left(\frac{q}{\beta}\right)^{n-1}(1-\beta^n E_n^2(x)), \tag{4.12}$$

hold for $\beta<q$, $n>1$ *and* $x\in(-(\beta^{1/2}+\beta^{-1/2})/2, (\beta^{1/2}+\beta^{-1/2})/2)$.

Acknowledgment. The authors acknowledge the referee's careful reading of the manuscript and helpful remarks.

REFERENCES

[1] R. ASKEY AND M. E. H. ISMAIL, *A generalization of the ultraspherical polynomials*, to appear.
[2] _____, *The Rogers q-ultraspherical polynomials*, in Approximation Theory III, E. W. Cheney, ed., Academic Press, New York, 1980, pp. 175–182.
[3] J. BUSTOZ, *Two parameter Turán inequalities*, J. Math. Anal. Appl., 79 (1981), pp. 71–79.
[4] J. BUSTOZ AND N. SAVAGE, *Inequalities for ultraspherical and Laguerre polynomials*, this Journal, 10 (1979), pp. 902–912.
[5] _____, *Inequalities for ultraspherical and Laguerre polynomials* II, this Journal, 11 (1980), pp. 876–884.
[6] G. FORSYTHE, *Second order determinants of Legendre polynomials*, Duke Math. J., 18 (1951), pp. 361–371.
[7] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[8] E. D. RAINVILLE, *Special Functions*, Macmillan, New York, 1960.
[9] P. SZÁSZ, *Inequalities concerning ultraspherical polynomials and Bessel functions*, Proc. Amer. Math. Soc., 1 (1950), pp. 256–267.
[10] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0015 $01.25/0

COHEN TYPE INEQUALITIES FOR JACOBI, LAGUERRE AND HERMITE EXPANSIONS*

C. MARKETT†

Abstract. Cohen's inequality was the first result on the way to the solution of Littlewood's conjecture. It is an estimate from below for the norm of a trigonometric polynomial in terms of the number of its nonzero coefficients. Inequalities of this type have been established in various other contexts, e.g., on compact groups or for Jacobi expansions. The purpose of this paper is to prove such inequalities for the classical orthogonal expansions in the appropriate weighted L^p spaces, here in terms of the highest coefficient. The results are best possible, apart from certain limiting values of the space parameter p.

1. Introduction and main results. By a "Cohen type inequality" we mean an estimate from below for the (weighted) L^1 norm of an (algebraic) polynomial p_n, as well as its generalization to the (weighted) L^p norms, $1<p\leq\infty$, which can be deduced from the L^1 estimate by means of corresponding Nikolskii type inequalities. Moreover, since one can often interpret a given polynomial p_n as kernel of a convolution operator, and the L^1 norm of p_n as the corresponding convolutor norm, also the extensions to L^p convolutor norms are usually called inequalities of Cohen type.

The above terminology comes from Cohen's contribution to the study of Littlewood's conjecture, which concerns the corresponding trigonometric problem. Littlewood [13] conjectured in 1948 that for any trigonometric polynomial $F_N(x)=\sum_{k=1}^{N} a_k e^{in_k x}$, where $0<n_1<\cdots<n_N$, $N\geq 2$, and $|a_k|\geq 1$ for $1\leq k\leq N$, there holds the estimate from below

$$\int_0^{2\pi}|F_N(x)|\,dx\geq C\log N, \tag{1.1}$$

where C is an absolute constant.

The first result in this direction is due to Cohen [3] who proved that $C(\log N/\log\log N)^{1/8}$ is a lower bound for the norm in (1.1). Several other authors made contributions to the problem by establishing larger bounds (cf. Fournier [7] for references), and recently, McGehee, Pigno and Smith [19] succeeded in confirming the Littlewood conjecture (1.1). Similar inequalities are also valid on compact groups (cf., e.g., [9], [19], [20]).

Giulini, Soardi and Travaglini [9] seem to be the first who formulated Cohen type inequalities in the sense of convolutor norm estimates and, recently, Dreseler and Soardi [4], [5] proved such estimates for Jacobi expansions (cf. Theorem A below). The purpose of this paper is to establish Cohen type inequalities for Laguerre and Hermite expansions. Our approach also admits a simpler proof of Theorem A. In each case, the result will be given in terms of convolutor norms (Theorems A, 1, 2) as well as in terms of weighted L^p norms of algebraic polynomials (Theorem 3).

We use the following notation. The Jacobi polynomials are given by

$$P_n^{\alpha,\beta}(x)=\sum_{k=0}^{n}\binom{n+\alpha}{n-k}\binom{n+\beta}{k}\left(\frac{x-1}{2}\right)^k\left(\frac{x+1}{2}\right)^{n-k},$$

*Received by the editors October 7, 1981.

†Lehrstuhl A für Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany.

where $\alpha, \beta > -1$, $x \in [-1,1]$, $n \in \mathbb{P} = \{0,1,2,\cdots\}$. If a function f belongs to one of the spaces

$$L^p_{w(\alpha,\beta)} = \begin{cases} \left\{ f;\ \left\{ \int_{-1}^{1} |f(x)|^p w^{\alpha,\beta}(x)dx \right\}^{1/p} < \infty \right\}, & 1 \le p < \infty, \\ \left\{ f;\ \operatorname*{ess\,sup}_{-1<x<1} |f(x)| < \infty \right\}, & p = \infty, \end{cases} \tag{1.2}$$

with $w^{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > -1$, its Jacobi expansion is given by $\sum_{k=0}^{\infty} \hat{f}(k) P_k^{\alpha,\beta}(x)$, where

(1.3)

$$\hat{f}(k) = h_k^{\alpha,\beta} \int_{-1}^{1} f(t) P_k^{\alpha,\beta}(t) w^{\alpha,\beta}(t)\, dt,$$

$$h_k^{\alpha,\beta} = \left\{ \int_{-1}^{1} \left[P_k^{\alpha,\beta}(x) \right]^2 w^{\alpha,\beta}(x)dx \right\}^{-1} = \frac{2k+\alpha+\beta+1}{2^{\alpha+\beta+1}} \frac{\Gamma(k+1)\Gamma(k+\alpha+\beta+1)}{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}.$$

In the following, $[X]$ always denotes the space of all bounded, linear operators from a space X into itself, endowed with the usual operator norm, $\| \cdot \|_{[X]}$.

THEOREM A. [5]. *Let $\alpha \ge \beta \ge -\frac{1}{2}$ and $1 \le p \le \infty$. For a given triangular matrix $\{c_{k,n}\}_{0 \le k \le n,\, n \in \mathbb{P}}$ of complex numbers with $|c_{n,n}| > 0$, let the operators $T_n^{\alpha,\beta}$ on $L^p_{w(\alpha,\beta)}$ be defined by $T_n^{\alpha,\beta} f = \sum_{k=0}^{n} c_{k,n} \hat{f}(k) P_k^{\alpha,\beta}$. There exists a positive constant C, independent of n, such that*

$$\| T_n^{\alpha,\beta} \|_{[L^p_{w(\alpha,\beta)}]} \ge C |c_{n,n}| d(\alpha, p, n), \tag{1.4}$$

$$d(\alpha,p,n) = \begin{cases} n^{(2\alpha+2)/p - (2\alpha+3)/2}, & 1 \le p < p_0(\alpha), \\ (\log(n+1))^{(2\alpha+1)/(4\alpha+4)}, & p = p_0(\alpha), \quad p = q_0(\alpha), \\ n^{(2\alpha+1)/2 - (2\alpha+2)/p}, & q_0(\alpha) < p \le \infty. \end{cases}$$

Here we have set $p_0(\alpha) = (4\alpha+4)/(2\alpha+3)$ and $q_0(\alpha) = (4\alpha+4)/(2\alpha+1)$.

In the particular case $\alpha = \beta \in (-\frac{1}{2}, \frac{1}{2})$ and $p = 1$ or $p = \infty$ the estimate (1.4) can be deduced from a result of Kal'nei [14],

$$\left\| T_n^{\alpha,\beta} \right\|_{[C[-1,1]]} \ge C n^{1/2} \sum_{k=0}^{n} |c_{k,n}| (k+1)^\alpha (n+1-k)^{-3/2-\alpha}.$$

It will be noted that Theorem A does not include the Chebyshev case $\alpha = \beta = -\frac{1}{2}$ for $p = 1$, i.e., the analogue of (1.1) for cosine polynomials. Besides the fact that the methods of proof of (1.1) and Theorem A are entirely different, the lower bound in (1.4) is given in terms of the highest frequency n whereas in (1.1) it is given in terms of the number of nonvanishing coefficients, $N \le n$. Thus, in the Jacobi case, one still has norm divergence if all $c_{k,n}$, $0 \le k \le n-1$, are zero; this obviously does not hold in the trigonometric case.

Laguerre expansions have been investigated mainly in the following two sets of weighted Lebesgue spaces, namely in the classical spaces (cf. [2], [22])

$$L^p_{u(\alpha)} = \left\{ f;\ \| f(x) u(x,\alpha) \|_{L^p(0,\infty)} < \infty, u(x,\alpha) = e^{-x/2} x^{\alpha/2} \right\} \tag{1.5}$$

for $\alpha > -\frac{2}{p}$ if $1 \le p < \infty$ and $\alpha \ge 0$ if $p = \infty$, as well as in the spaces

$$L^p_{w(\alpha)} = \begin{cases} \left\{ f: \left\{ \int_0^\infty |f(x)e^{-x/2}|^p x^\alpha \, dx \right\}^{1/p} < \infty \right\}, & 1 \le p < \infty, \\ \left\{ f; \operatorname*{ess\,sup}_{x>0} |f(x)e^{-x/2}| < \infty \right\}, & p = \infty, \end{cases} \tag{1.6}$$

for $\alpha > -1$; the latter have recently been shown by Görlich and Markett [12] to be particularly suited for introducing a convolution structure. Both types of norm will be considered here. A typical result for the spaces $L^p_{w(\alpha)}$, $\alpha \ge 0$, will be that the lower bound for the operator T_n^α obtained in (1.8) below is the same as the one for the Jacobi case. For the particular case of the partial sum operators a similar effect has been observed in [12].

Denoting the Laguerre polynomials for $n \in \mathbb{P}$ and $\alpha > -1$ by

$$L_n^\alpha(x) = \begin{cases} (n!)^{-1} e^x x^{-\alpha} (d/dx)^n (e^{-x} x^{\alpha+n}), & x > 0, \\ A_n^\alpha = \binom{n+\alpha}{n}, & x = 0, \end{cases}$$

the Laguerre expansion of a function f is given by $\sum_{k=0}^\infty \hat{f}(k) L_k^\alpha(x)$, where

$$\hat{f}(k) = \hat{f}(k,\alpha) = h_k^\alpha \int_0^\infty f(t) L_k^\alpha(t) e^{-t} t^\alpha \, dt, \tag{1.7}$$

$$h_k^\alpha = \left\{ \int_0^\infty [L_k^\alpha(x)]^2 e^{-x} x^\alpha \, dx \right\}^{-1} = \Gamma(k+1)/\Gamma(k+\alpha+1),$$

as far as these expressions make sense.

Our first main result is

THEOREM 1. *Let $\alpha \ge 0$ and $1 \le p \le \infty$. For a given triangular matrix $\{c_{k,n}\}_{0 \le k \le n,\, n \in \mathbb{P}}$ of complex numbers with $|c_{n,n}| > 0$, let the operators T_n^α on $L^p_{w(\alpha)}$ or $L^p_{u(\alpha)}$ be defined by $T_n^\alpha f = \sum_{k=0}^n c_{k,n} \hat{f}(k) L_k^\alpha$. There exist positive constants C_1, C_2, independent of n, such that (with $d(\alpha,p,n)$ as in Theorem A)*

$$\|T_n^\alpha\|_{[L^p_{w(\alpha)}]} \ge C_1 |c_{n,n}| d(\alpha,p,n), \tag{1.8}$$

$$\|T_n^\alpha\|_{[L^p_{u(\alpha)}]} \ge C_2 |c_{n,n}| \begin{cases} n^{2/p-3/2}, & 1 \le p < \frac{4}{3}, \\ (\log(n+1))^{1/4}, & p = \frac{4}{3},\ p = 4, \\ n^{1/2-2/p}, & 4 < p \le \infty. \end{cases} \tag{1.9}$$

Finally let the Hermite polynomials be given by

$$H_n(x) = (-1)^n e^{x^2} \left(\frac{d}{dx} \right)^n e^{-x^2} \qquad (x \in \mathbb{R}, n \in \mathbb{P}).$$

To a function f belonging to one of the spaces (cf. [2], [22])

$$L^p_{u(H)} = \left\{ f; \| f(x)u(x) \|_{L^p(-\infty,\infty)} < \infty, u(x) = e^{-x^2/2} \right\}, \qquad 1 \le p \le \infty, \tag{1.10}$$

one associates the Hermite expansion $\sum_{k=0}^\infty \hat{f}(k) H_k(x)$, where

$$\hat{f}(k) = h_k^H \int_{-\infty}^\infty f(t) H_k(t) e^{-t^2} \, dt, \tag{1.11}$$

$$h_k^H = \left\{ \int_{-\infty}^\infty [H_k(x)]^2 e^{-x^2} \, dx \right\}^{-1} = \left(\sqrt{\pi}\, 2^k k! \right)^{-1}.$$

THEOREM 2. *For a given triangular matrix* $\{c_{k,n}\}_{0\le k\le n,\, n\in\mathbb{P}}$ *of complex numbers with* $|c_{n,n}|>0$, *let the operators* T_n^H *be defined on* $L^p_{u(H)}$, $1\le p\le\infty$, *by* $T_n^H f=\sum_{k=0}^n c_{k,n}\hat{f}(k)H_k$. *There exists a positive constant* C, *independent of* n, *such that*

$$\|T_n^H\|_{[L^p_{u(H)}]}\ge C|c_{n,n}|\begin{cases} n^{2/(3p)-1/2}, & 1\le p<\frac{4}{3},\\ (\log(n+1))^{1/4}, & p=\frac{4}{3},\ p=4,\\ n^{1/6-2/(3p)}, & 4<p\le\infty. \end{cases}$$

Remarks. i) Theorem A is best possible as far as the exponents of n in case $1\le p<p_0(\alpha)$ and $q_0(\alpha)<p\le\infty$ are concerned, as may be seen by choosing for $T_n^{\alpha,\beta}$ the partial sum operators of the Jacobi expansion ($c_{k,n}=1$, $0\le k\le n$). In the limiting cases $p=p_0(\alpha)$ and $p=q_0(\alpha)$, however, the question as to whether the exponent of $\log(n+1)$ is sharp cannot be decided in this way (cf. Lemma 3 below).

ii) A similar remark applies to Theorems 1 and 2, as follows from known lower bounds for the Laguerre and Hermite partial sums (cf. [12], [15], [17]).

iii) When applied to Cesàro summation operators $(C,\delta)_n$, i.e., $c_{k,n}=A_{n-k}^{\delta}/A_n^{\delta}$, $c_{n,n}\sim n^{-\delta}$ $(n\to\infty)$, Theorems A, 1 and 2 also yield sharp results with possible exception of the limiting cases (cf. [10], [12], [15], [17]).

With some additional arguments one can derive from the above theorems:

THEOREM 3. *Let* $\{c_{k,n}\}_{0\le k\le n,\, n\in\mathbb{P}}$ *be a triangular matrix of complex numbers with* $|c_{n,n}|>0$. *Then for* $1\le p\le\infty$, $n\in\mathbb{N}$ *one has*

a) *in the Jacobi case for* $\alpha\ge\beta\ge-\frac{1}{2}$, $\alpha>-\frac{1}{2}$,

$$\left\|\sum_{k=0}^n c_{k,n}P_k^{\alpha,\beta}(x)\right\|_{L^p_{w(\alpha,\beta)}}\ge C|c_{n,n}|n^{-1/2};\tag{1.12}$$

b) *in the Laguerre case for* $\alpha,\beta>-1$,

$$\left\|\sum_{k=0}^n c_{k,n}L_k^{\beta}(x)\right\|_{L^p_{u(\alpha)}}\ge C|c_{n,n}|n^{(\alpha+1)/p-1/2}\tag{1.13}$$

and, for $\beta>-1$, $\alpha>-\frac{2}{p}$ *if* $1\le p<\infty$ *and* $\alpha\ge0$ *if* $p=\infty$,

$$\left\|\sum_{k=0}^n c_{k,n}L_k^{\beta}(x)\right\|_{L^p_{w(\alpha)}}\ge C|c_{n,n}|n^{1/p+(\alpha-1)/2};\tag{1.14}$$

c) *in the Hermite case*,

$$\left\|\sum_{k=0}^n c_{k,n}H_k(x)\right\|_{L^p_{u(H)}}\ge C|c_{n,n}|\left(\sqrt{\pi}\,2^n n!\right)^{1/2}n^{1/(2p)-1/4}.\tag{1.15}$$

Moreover, these bounds are best possible in the sense that triangular matrices $\{c_{k,n}\}_{0\le k\le n,\, n\in\mathbb{P}}$ *can be given for which the estimates are sharp.*

We proceed as follows. In §2 we supply some properties and norm estimates of the orthogonal polynomials required. The proofs of Theorems A, 1 and 2 will be given in §3, and the proof of Theorem 3 follows in §4. Principally, the proofs of Theorem A, 1 and 2 have the common feature that they start with the upper p intervals, i.e., for $p\ge q_0(\alpha)$ in Theorem A and (1.8) and for $p\ge4$ in (1.9) and Theorem 2, respectively, and the assertions for the lower p intervals then follow by duality.

For the proof of Theorem A, Dreseler and Soardi used the test functions

$$f_n^{\alpha,\beta}(\cos\phi)=\sin(n+m)\phi\sin^{m-(2\alpha+1)}\frac{\phi}{2}\cos^{m-(2\beta+1)}\frac{\phi}{2},$$

$0\le\phi\le\pi$, denoting by m the smallest odd integer for which $m\ge 2\alpha+1$. Here the crucial step of the proof was to show that $(f_n^{\alpha,\beta})\hat{}\,(k)=0$ for $0\le k\le n-1$ by representing $P_n^{\alpha,\beta}(\cos\phi)$ and $\sin(n+m)\phi\sin^m\phi$ as cosine polynomials and to find the exact rate of growth of $(f_n^{\alpha,\beta})\hat{}\,(n)$. We will give another proof of Theorem A by means of the test functions

$$(1.16)\qquad g_n^{\alpha,\beta,j}(x)=P_n^{\alpha+j,\beta+j}(x)(1-x^2)^j,\qquad j\in\mathbb{P},\quad j>\alpha+\frac{1}{2}-\frac{2\alpha+2}{p},$$

the Jacobi expansion of which immediately yields that $(g_n^{\alpha,\beta,j})\hat{}\,(k)=0$ for $0\le k\le n-1$ and $(g_n^{\alpha,\beta,j})\hat{}\,(n)\ge C>0$ (see (2.8)).

A similar construction may be made in the Laguerre case. For the proof of (1.8) we use the test polynomials

$$(1.17)\qquad g_n^{\alpha,j}(x)=L_n^{\alpha+j}(x)x^j-\left(\frac{(n+1)(n+2)}{(n+\alpha+j+1)(n+\alpha+j+2)}\right)^{1/2}L_{n+2}^{\alpha+j}(x)x^j,$$

with $j\in\mathbb{P}$, $j>\alpha-\frac{1}{2}-(2\alpha+2)/p$, the first term of which is of the same type as the function in (1.16), whereas the second term does not affect $\mathrm{T}_n^{\alpha}g_n^{\alpha,j}$ but serves for smoothing the norm behaviour of $g_n^{\alpha,j}$. Using the Laguerre functions

$$(1.18)\qquad \mathfrak{L}_n^{\alpha}(x)=\left(\frac{n!}{\Gamma(n+\alpha+1)}\right)^{1/2}L_n^{\alpha}(x)e^{-x/2}x^{\alpha/2}\qquad(\alpha>-1,\ n\in\mathbb{P},\ x\ge 0),$$

the $g_n^{\alpha,j}$ may also be written as

$$(1.19)\qquad g_n^{\alpha,j}(x)=\left(\frac{\Gamma(n+\alpha+j+1)}{n!}\right)^{1/2}\{\mathfrak{L}_n^{\alpha+j}(x)-\mathfrak{L}_{n+2}^{\alpha+j}(x)\}e^{x/2}x^{(j-\alpha)/2}.$$

In the case of (1.9), it suffices to use the test functions $g_n^{\alpha,0}$.

Similarly, in the Hermite case the test functions

$$(1.20)\qquad \begin{aligned} g_n^{H}(x)&=H_n(x)-[4(n+1)(n+2)]^{-1/2}H_{n+2}(x)\\ &=(\pi^{1/2}2^n n!)^{1/2}\{\mathcal{H}_n(x)-\mathcal{H}_{n+2}(x)\}e^{x^2/2}\end{aligned}$$

will be used to establish Theorem 2. Here $\mathcal{H}_n(x)$ denotes the Hermite function

$$(1.21)\qquad \mathcal{H}_n(x)=(\pi^{1/2}2^n n!)^{-1/2}H_n(x)e^{-x^2/2}.$$

2. Preliminaries. Let us recall some properties of the Jacobi polynomials for $\alpha,\beta>-1$.

$$(2.1)\qquad P_n^{\alpha,\beta}(x)=(-1)^n P_n^{\beta,\alpha}(-x)\qquad(-1\le x\le 1,\ n\in\mathbb{P}),$$

$$(2.2)\qquad \|P_n^{\alpha,\beta}(x)(1-x)^{\gamma}\|_{L^p(0,1)}\sim\begin{cases} n^{-1/2}, & \gamma>\dfrac{\alpha}{2}+\dfrac{1}{4}-\dfrac{1}{p},\\ n^{-1/2}(\log n)^{1/p}, & \gamma=\dfrac{\alpha}{2}+\dfrac{1}{4}-\dfrac{1}{p},\\ n^{\alpha-2\gamma-2/p}, & \gamma<\dfrac{\alpha}{2}+\dfrac{1}{4}-\dfrac{1}{p}\end{cases}\quad(n\to\infty),$$

where $\gamma>-\frac{1}{p}$ if $1\le p<\infty$ and $\gamma\ge 0$ if $p=\infty$. Here, "$a_n\sim b_n$" stands for $a_n=O(b_n)$ and $b_n=O(a_n)$ as $n\to\infty$. For $p=1$ see [23,(7.34.1)]. Relation (2.2) was formulated for

general $p>0$ in [23, Problem 91] without proof and can be verified by using the bounds for Jacobi polynomials given in [23, (7.32.5)], [21, (2.4), (2.5)].

In the following, the Jacobi expansion of the test functions $g_n^{\alpha,\beta,j}$ in (1.16) will be derived. We start from the formula [23, (9.4.3)],

$$(2.3)\quad P_n^{\alpha+\mu,\beta}(x)=\frac{\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+\mu+1)}\sum_{k=0}^{n}A_{n-k}^{\mu-1}\frac{\Gamma(n+k+\alpha+\beta+\mu+1)}{\Gamma(n+k+\alpha+\beta+2)}\cdot(2k+\alpha+\beta+1)\frac{\Gamma(k+\alpha+\beta+1)}{\Gamma(k+\beta+1)}P_k^{\alpha,\beta}(x),$$

which can be inverted to give (cf. [1, (13), (14)])

$$(2.4)\quad (1-x)^{\mu}P_n^{\alpha+\mu,\beta}(x)=\frac{\Gamma(n+\alpha+\mu+1)}{\Gamma(n+1)}2^{\mu}\sum_{k=0}^{\infty}A_k^{-\mu-1}\frac{\Gamma(2n+k+\alpha+\beta+1)}{\Gamma(2n+k+\alpha+\beta+\mu+2)}\cdot(2n+2k+\alpha+\beta+1)\frac{\Gamma(n+k+1)}{\Gamma(n+k+\alpha+1)}P_{n+k}^{\alpha,\beta}(x).$$

For $\mu=j\in\mathbb{P}$, the series in (2.4) terminates, so that

$$(2.5)\qquad (1-x)^{j}P_n^{\alpha+j,\beta}(x)=\sum_{k=0}^{j}a_{k,j}(\alpha,\beta,n)P_{n+k}^{\alpha,\beta}(x)$$

with certain coefficients $a_{k,j}$, the first and last of which read

$$(2.6)\qquad a_{0,j}(\alpha,\beta,n)=2^{j}\frac{\Gamma(n+\alpha+j+1)}{\Gamma(n+\alpha+1)}\frac{\Gamma(2n+\alpha+\beta+2)}{\Gamma(2n+\alpha+\beta+j+2)},$$

$$a_{j,j}(\alpha,\beta,n)=(-2)^{j}\frac{\Gamma(n+j+1)}{\Gamma(n+1)}\frac{\Gamma(2n+j+\alpha+\beta+1)}{\Gamma(2n+2j+\alpha+\beta+1)}.$$

In view of (2.1) this also implies

$$(2.7)\qquad (1+x)^{j}P_n^{\alpha,\beta+j}(x)=\sum_{k=0}^{j}(-1)^{k}a_{k,j}(\beta,\alpha,n)P_{n+k}^{\alpha,\beta}(x).$$

Combining (2.5) to (2.7), it follows that

(2.8)

$$(1-x^2)^{j}P_n^{\alpha+j,\beta+j}(x)=\sum_{k=0}^{2j}b_{k,j}(\alpha,\beta,n)P_{n+k}^{\alpha,\beta}(x),$$

$$b_{0,j}(\alpha,\beta,n)=a_{0,j}(\alpha,\beta+j,n)a_{0,j}(\beta,\alpha,n)=4^{j}\frac{\Gamma(n+\alpha+j+1)}{\Gamma(n+\alpha+1)}\frac{\Gamma(n+\beta+j+1)}{\Gamma(n+\beta+1)}\frac{\Gamma(2n+\alpha+\beta+2)}{\Gamma(2n+\alpha+\beta+2j+2)},$$

$$b_{2j,j}(\alpha,\beta,n)=a_{j,j}(\alpha,\beta+j,n)(-1)^{j}a_{j,j}(\beta,\alpha,n+j)=(-4)^{j}\frac{\Gamma(n+2j+1)}{\Gamma(n+1)}\frac{\Gamma(2n+2j+\alpha+\beta+1)}{\Gamma(2n+4j+\alpha+\beta+1)}.$$

In the Laguerre case, the following properties will be used [22]. For $\nu=\nu(n,\alpha)=4n+2\alpha+2$, $n\in\mathbb{N}$, $\alpha>-1$, there exist positive constants C and γ, such that

$$(2.9)\qquad |\mathcal{L}_{n+1}^{\alpha}(x)-\mathcal{L}_{n-1}^{\alpha}(x)|\le C\begin{cases}\nu^{\alpha/2-1}x^{\alpha/2}, & 0\le x\le\dfrac{1}{\nu},\\ \nu^{-3/4}x^{1/4}, & \dfrac{1}{\nu}<x\le\dfrac{\nu}{2},\\ \nu^{-3/4}\left(\nu^{1/3}+|x-\nu|\right)^{1/4}, & \dfrac{\nu}{2}<x\le\dfrac{3\nu}{2},\\ e^{-\gamma x}, & \dfrac{3\nu}{2}<x.\end{cases}$$

$$(2.10)\qquad \mathcal{L}_{n+1}^{\alpha}(x)-\mathcal{L}_{n-1}^{\alpha}(x)=\frac{2^{5/2}(-1)^{n-1}x^{1/4}}{\pi^{1/2}\nu^{3/4}}\left(\sin g+O\left[\frac{x}{\nu}+\frac{1}{x}\right]\right),\qquad 1\le x\le\frac{\nu}{2},$$

where $g=g(\nu,x^{1/2})=-\frac{\nu}{2}\cos^{-1}((\frac{x}{\nu})^{1/2})+\frac{1}{2}(x(\nu-x))^{1/2}+\frac{\pi}{4}$ and the O-term holds uniformly with respect to $x\in[1,\nu/2]$ if $n\to\infty$.

LEMMA 1. *For $\alpha\ge 0$ and $q_0(\alpha)=(4\alpha+4)/(2\alpha+1)$,*

$$(2.11)\qquad \|L_n^{\alpha}(x)\|_{L^p_{w(\alpha)}}\sim\begin{cases}n^{\alpha/2-1/4}(\log n)^{1/q_0}, & p=q_0,\\ n^{\alpha-(\alpha+1)/p}, & q_0<p\le\infty\end{cases}\quad(n\to\infty).$$

For $\alpha>-1$, $\beta>-\min\{\alpha/2,1/4\}-1/p$ if $1\le p<\infty$ and $\beta\ge-\min\{\alpha/2,1/4\}$ if $p=\infty$,

$$(2.12)\qquad \left\|\left[\mathcal{L}_{n+1}^{\alpha}(x)-\mathcal{L}_{n-1}^{\alpha}(x)\right]x^{\beta}\right\|_{L^p(0,\infty)}\sim n^{\beta-1/2+1/p}\qquad(n\to\infty).$$

Proof. Concerning (2.11), use [16, Lemma 1] with α,β replaced by $\frac{2\alpha}{p}$ and $\alpha-\frac{2\alpha}{p}$, respectively. For $\beta=0$, relation (2.12) has also been proved in [16, Lemma 1]. In the general case, we follow the same lines as there. The upper estimate is obtained by applying the various bounds of (2.9) to the integrand. For the lower estimate, we restrict the interval of integration to $[1,b\nu/2]$ and use the asymptotic expansion (2.10). Here we suppose that ν is large enough in order to apply [22, Lemma 15], whereas $b\in(0,1)$ has to be small enough, such that the principal term exceeds the two other terms. □

In order to find the Laguerre expansion for the test functions $g_n^{\alpha,j}$ of (1.16), we start with formula

$$(2.13)\qquad L_n^{\alpha+\mu}(x)=\sum_{k=0}^{n}A_{n-k}^{\mu-1}L_k^{\alpha}(x),$$

which can be inverted to give

$$(2.14)\qquad x^{\mu}L_n^{\alpha+\mu}(x)=\frac{\Gamma(n+\alpha+\mu+1)}{\Gamma(n+1)}\sum_{k=0}^{\infty}A_k^{-\mu-1}\frac{\Gamma(n+k+1)}{\Gamma(n+k+\alpha+1)}L_{n+k}^{\alpha}(x)$$

(cf. [1,(9)] or [6,§6.15.1(1)] for a more general formula on confluent hypergeometric functions). For $\mu=j\in\mathbb{P}$ the series terminates as in (2.4), (2.5), so that by definition (1.17)

$$(2.15)\qquad g_n^{\alpha,j}(x)=\sum_{k=0}^{j+2}b_{k,j}(\alpha,n)L_{n+k}^{\alpha}(x),$$

$$b_{0,j}(\alpha,n)=\frac{\Gamma(n+\alpha+j+1)}{\Gamma(n+\alpha+1)},$$

$$b_{j+2,j}(\alpha,n)=(-1)^{j+1}\left(\frac{(n+1)(n+2)}{(n+\alpha+j+1)(n+\alpha+j+2)}\right)^{1/2}\frac{\Gamma(n+j+3)}{\Gamma(n+3)}.$$

For any algebraic polynomial of degree not exceeding $n\in\mathbb{N}$, $p_n\in P_n$, the following general inequality holds. In particular, for $\gamma=\alpha/p$, it is of Nikolskii-type.

LEMMA 2. *Let $1\le p\le\infty$ and let parameters α,γ be given such that $\gamma>-1/p$ if $1\le p<\infty$, and $\gamma\ge 0$ if $p=\infty$, as well as $\alpha\ge\gamma$ if $p=1$, and $\alpha>\gamma+1/p-1$ if $1<p\le\infty$. There exists a positive constant C such that for each $p_n\in P_n$, $n\in\mathbb{N}$,*

$$\|p\|_{L^1_{w(\alpha)}}\le Cn^{\alpha-\gamma+1-1/p}\|p_n(x)e^{-x/2}x^{\gamma}\|_{L^p(0,\infty)}.$$

Proof. For $\nu=\nu(n,\gamma)=4n+2\gamma+2$, we write

$$\|p_n\|_{L^1_{w(\alpha)}}=\left\{\int_0^{3\nu/2}+\int_{3\nu/2}^{\infty}\right\}|p_n(x)e^{-x/2}x^{\alpha}|\,dx=I_1+I_2,$$

say. Hölder's inequality yields

$$\begin{aligned}I_1&\le\|p_n(x)e^{-x/2}x^{\gamma}\|_{L^p(0,3\nu/2)}\|x^{\alpha-\gamma}\|_{L^q(0,3\nu/2)}\\&\le C\|p_n(x)e^{-x/2}x^{\gamma}\|_{L^p(0,\infty)}n^{\alpha-\gamma+1-1/p},\end{aligned}$$

and, in view of the exponential decrease of $|\mathfrak{L}_k^{\gamma}(x)|$, $0\le k\le n$, for $x>3\nu(n,\gamma)/2$ [22, (2.5)], one has

$$\begin{aligned}I_2&=\int_{3\nu/2}^{\infty}\left|\sum_{k=0}^{n}\hat{p_n}(k,\gamma)L_k^{\gamma}(x)e^{-x/2}x^{\alpha}\right|dx\\&\le\sum_{k=0}^{n}h_k^{\gamma}\left|\int_0^{\infty}p_n(t)L_k^{\gamma}(t)e^{-t}t^{\gamma}\,dt\right|\int_{3\nu/2}^{\infty}\left|L_k^{\gamma}(x)e^{-x/2}x^{\alpha}\right|dx\\&\le\|p_n(t)e^{-t/2}t^{\gamma}\|_{L^p(0,\infty)}\\&\quad\cdot\sum_{k=0}^{n}h_k^{\gamma}\|L_k^{\gamma}(t)e^{-t/2}\|_{L^q(0,\infty)}\int_{3\nu/2}^{\infty}\left|\mathfrak{L}_k^{\gamma}(x)x^{\alpha-\gamma/2}\right|dx\\&\le C\|p_n(t)e^{-t/2}t^{\gamma}\|_{L^p(0,\infty)}.\end{aligned}$$

The norm behaviour of the Hermite functions (1.12) and their differences can be taken over from [17, Lemma 1].

$$\|\mathcal{H}_n(x)\|_{L^p(-\infty,\infty)}\sim\begin{cases}n^{-1/8}(\log n)^{1/4}, & p=4,\\ n^{-1/(6p)-1/12}, & 4<p\le\infty\end{cases}\quad(n\to\infty),\tag{2.16}$$

$$\|\mathcal{H}_{n+1}(x)-\mathcal{H}_{n-1}(x)\|_{L^p(-\infty,\infty)}\sim n^{1/(2p)-1/4}\qquad(n\to\infty).\tag{2.17}$$

3. Proofs of Theorems A, 1 and 2, Lemma 3.

Proof of Theorem A. Let $\alpha\ge\beta\ge-\frac{1}{2}$, $\alpha>-\frac{1}{2}$. By duality it suffices to assume that $q_0(\alpha)\le p\le\infty$. Then Theorem A will be established by applying the operators $T_n^{\alpha,\beta}$ to the test functions $g_n^{\alpha,\beta,j}$ for some $j\in\mathbb{P}$, $j>\alpha+\frac{1}{2}-(2\alpha+2)/p$. In view of (2.8), they satisfy $(g_n^{\alpha,\beta,j})\hat{}\,(k)=0$ if $0\le k\le n-1$ and $=b_{0,j}\ge C$ if $k=n$. Moreover, by (2.1) and

Lemma 1, one has for $n\in\mathbb{N}$

$$\|g_n^{\alpha,\beta,j}\|_{L^p_{w(\alpha,\beta)}}\le C\Big\{\|P_n^{\alpha+j,\beta+j}(x)(1-x)^{j+\alpha/p}\|_{L^p(0,1)} + \|P_n^{\beta+j,\alpha+j}(x)(1-x)^{j+\beta/p}\|_{L^p(0,1)}\Big\} \le Cn^{-1/2} \tag{3.1}$$

since $j>\alpha+\frac{1}{2}-(2\alpha+2)/p\ge\beta+\frac{1}{2}-(2\beta+2)/p$. Thus,

$$\|T_n^{\alpha,\beta}\|_{[L^p_{w(\alpha,\beta)}]}\ge\Big[\|g_n^{\alpha,\beta,j}\|_{L^p_{w(\alpha,\beta)}}\Big]^{-1}\|T_n^{\alpha,\beta}g_n^{\alpha,\beta,j}\|_{L^p_{w(\alpha,\beta)}} \ge C|c_{n,n}|n^{1/2}\|P_n^{\alpha,\beta}\|_{L^p_{w(\alpha,\beta)}}. \tag{3.2}$$

An application of Lemma 1 with $\gamma=\frac{\alpha}{p}$ then yields (1.4). □

Proof of Theorem 1. By duality, we may again restrict p to $q_0(\alpha)\le p\le\infty$ in (1.8) and to $4\le p\le\infty$ in (1.9). Concerning (1.8), we apply the operators T_n^α to the test functions $g_n^{\alpha,j}$ of (1.17) for some $j\in\mathbb{P}$, $j>\alpha-\frac{1}{2}-(2\alpha+2)/p$. Their Laguerre expansions (2.15) give that $(g_n^{\alpha,j})\hat{}\,(k)=0$ if $0\le k\le n-1$ and $(g_n^{\alpha,j})\hat{}\,(n)=b_{0,j}(\alpha,n)\ge Cn^j$. Furthermore, in view of (1.6), (1.19) and (2.12) of Lemma 1, with α,β replaced by $\alpha+j$ and $(j-\alpha)/2+\frac{\alpha}{p}$, respectively, one has

$$\|g_n^{\alpha,j}(x)\|_{L^p_{w(\alpha)}}\le Cn^{(\alpha+j)/2}\Big\|\Big[\mathfrak{L}_n^{\alpha+j}(x)-\mathfrak{L}_{n+2}^{\alpha+j}(x)\Big]x^{(j-\alpha)/2+\alpha/p}\Big\|_{L^p(0,\infty)} \le Cn^{j-1/2+(\alpha+1)/p}\qquad(n\in\mathbb{N}). \tag{3.3}$$

This implies

$$\|T_n^\alpha\|_{[L^p_{w(\alpha)}]}\ge\Big[\|g_n^{\alpha,j}\|_{L^p_{w(\alpha)}}\Big]^{-1}\|T_n^\alpha g_n^{\alpha,j}\|_{L^p_{w(\alpha)}} \ge C|c_{n,n}|n^jn^{-j+1/2-(\alpha+1)/p}\|L_n^\alpha\|_{L^p_{w(\alpha)}}. \tag{3.4}$$

The assertion now follows by Lemma 1 and (2.11).

Concerning (1.9), we use (1.5), (1.19), (2.12) with $\beta=0$ and a lower estimate of the $L^p_{u(\alpha)}$ norms of the L_n^α given in [16, Lemma 1], to deduce

$$\begin{aligned}\|T_n^\alpha\|_{[L^p_{u(\alpha)}]}&\ge\Big[\|g_n^{\alpha,0}\|_{L^p_{u(\alpha)}}\Big]^{-1}\|T_n^\alpha g_n^{\alpha,0}\|_{L^p_{u(\alpha)}}\\ &\ge C|c_{n,n}|\Big[n^{\alpha/2}\|\mathfrak{L}_n^\alpha(x)-\mathfrak{L}_{n+2}^\alpha(x)\|_{L^p(0,\infty)}\Big]^{-1}\|L_n^\alpha\|_{L^p_{u(\alpha)}}\\ &\ge C|c_{n,n}|n^{-\alpha/2+1/2-1/p}\begin{cases}n^{\alpha/2-1/2+1/p}(\log(n+1))^{1/p}, & p=4,\\ n^{\alpha/2-1/p}, & 4<p\le\infty.\end{cases}\end{aligned}$$

□

Proof of Theorem 2. Applying the operators T_n^H to the test functions g_n^H as given in (1.20), Theorem 2 is proved analogously to the estimate (1.9), in view of (1.10), (1.21), (2.16) and (2.17). □

We conclude this section with a lemma which gives rise to the assumption that, in contrast to the cases $p<p_0(\alpha)$ or $p>q_0(\alpha)$, the exponent in Theorem A may possibly be enlarged in the limiting cases (cf. Remark i). The proof of this lemma will be dual to the one of Theorem A in the following sense. Whereas the test functions $g_n^{\alpha,\beta,j}$ of (1.16), which were used in proving Theorem A for the upper interval $p\in[q_0(\alpha),\infty]$, may be considered as *differences of order j* of Jacobi polynomials $P_n^{\alpha,\beta}$ with respect to both the parameters α and β (cf. (2.4) to (2.8)), the test functions to be used for Lemma 3 will

work in the lower interval $p\in[1,p_0(\alpha)]$, and they will be *sums of order j* of Jacobi polynomials.

LEMMA 3. *Let* $\alpha\geq\beta\geq-\frac{1}{2}$, $\alpha>-\frac{1}{2}$ *and let* $p_0(\alpha)$ *and* $q_0(\alpha)$ *be defined as in Theorem A. For each* $n\in\mathbb{N}$ *the Jacobi partial sums satisfy*

$$\|S_n^{\alpha,\beta}\|_{[L^p_{w(\alpha,\beta)}]}\geq C\begin{cases} n^{(2\alpha+2)/p-(2\alpha+3)/2}, & 1\leq p<p_0,\\ (\log(n+1))^{(2\alpha+3)/(4\alpha+4)}, & p=p_0, p=q_0,\\ n^{(2\alpha+1)/2-(2\alpha+2)/p}, & q_0<p\leq\infty. \end{cases}$$

Proof. Applying $S_n^{\alpha,\beta}$ to the test functions

$$P_{2n}^{\alpha+j+1,\beta}(x),\qquad j=\left[\frac{2\alpha+2}{p}-\frac{2\alpha+3}{2}\right]+1 \tag{3.5}$$

([b] denoting the greatest integer less than or equal to b) and using (2.3) with $\mu=j+1$, one obtains, after a j-fold partial summation,

$$S_n^{\alpha,\beta}\left(P_{2n}^{\alpha+j+1,\beta};x\right)$$

$$=\frac{\Gamma(2n+\beta+1)}{\Gamma(2n+\alpha+\beta+j+2)}\sum_{k=0}^{n}A_{2n-k}^{j}\frac{\Gamma(2n+k+\alpha+\beta+j+2)}{\Gamma(2n+k+\alpha+\beta+2)}(2k+\alpha+\beta+1)$$
$$\cdot\frac{\Gamma(k+\alpha+\beta+1)}{\Gamma(k+\beta+1)}P_k^{\alpha,\beta}(x)$$

$$=\frac{\Gamma(2n+\beta+1)}{\Gamma(2n+\alpha+\beta+j+2)}\sum_{m=0}^{j}A_{n+m}^{j-m}\frac{\Gamma(3n-m+\alpha+\beta+j+2)}{\Gamma(3n+\alpha+\beta+2)}$$
$$\cdot\frac{\Gamma(n+\alpha+\beta+2)}{\Gamma(n-m+\beta+1)}P_{n-m}^{\alpha+m+1,\beta}(x).$$

Taking norms on both sides, the term for $m=0$ on the right-hand side turns out to be the principal one, so that

$$\|S_n^{\alpha,\beta}\|_{[L^p_{w(\alpha,\beta)}]}\geq Cn^j\|P_n^{\alpha+1,\beta}\|_{L^p_{w(\alpha,\beta)}}/\|P_{2n}^{\alpha+j+1,\beta}\|_{L^p_{w(\alpha,\beta)}}.$$

Since $j>(2\alpha+2)/p-(2\alpha+3)/2$, Lemma 1 yields the assertion for $1\leq p\leq p_0(\alpha)$. The assertion for $q_0(\alpha)\leq p\leq\infty$ then follows by duality. □

Analogous results also hold in the Laguerre case (cf. [11]).

4. Proof of Theorem 3. a) *Jacobi case.* Let $\alpha\geq\beta\geq-\frac{1}{2}$, $\alpha>-\frac{1}{2}$ and, first of all, $p=1$. For a triangular matrix $\{d_{k,n}\}_{0\leq k\leq n,\,n\in\mathbb{P}}$ with $|d_{n,n}|>0$, we define a convolution operator on $L^1_{w(\alpha,\beta)}$ by

$$T_n^{\alpha,\beta}f=\sum_{k=0}^{n}d_{k,n}\hat{f}(k)P_k^{\alpha,\beta}=f*k_n^{\alpha,\beta},$$

with the kernel function

$$k_n^{\alpha,\beta}(x)=\sum_{k=0}^{n}d_{k,n}h_k^{\alpha,\beta}P_k^{\alpha,\beta}(1)P_k^{\alpha,\beta}(x)\qquad(x\in[-1,1]).$$

Setting in particular $d_{k,n}=c_{k,n}\{h_k^{\alpha,\beta}P_k^{\alpha,\beta}(1)\}^{-1}$, where $h_k^{\alpha,\beta}\sim k$ and $P_k^{\alpha,\beta}(1)=\binom{k+\alpha}{k}\sim k^\alpha$ as $k\to\infty$, it follows by the convolution theorem for Jacobi expansions [8] and Theorem A that

$$\left\|\sum_{k=0}^{n} c_{k,n}P_k^{\alpha,\beta}\right\|_{L^1_{w(\alpha,\beta)}}=\left\|k_n^{\alpha,\beta}\right\|_{L^1_{w(\alpha,\beta)}}=\left\|T_n^{\alpha,\beta}\right\|_{[L^1_{w(\alpha,\beta)}]}$$
$$\geq C|d_{n,n}|n^{\alpha+1/2}\geq C|c_{n,n}|n^{-1/2}\qquad(n\in\mathbb{N}).$$

For $p>1$, we apply the trivial norm inequality

$$\|p_n\|_{L^1_{w(\alpha,\beta)}}\leq C\|p_n\|_{L^p_{w(\alpha,\beta)}}\qquad(p_n\in P_n,\ 1\leq p\leq\infty),$$

to derive (1.12).

In order to show that the estimate (1.12) is best possible for each $1\leq p\leq\infty$, we can choose the test functions (1.16) with n replaced by $n-2j$,

$$g_{n-2j}^{\alpha,\beta,j}(x)=P_{n-2j}^{\alpha+j,\beta+j}(x)(1-x^2)^j,$$

where $n\geq 2j+1$ and $j>\max\{\alpha+\frac{1}{2}-(2\alpha+2)/p,\ \beta+\frac{1}{2}-(2\beta+2)/p\}$. Indeed, in view of (2.8), an application of (1.12) gives

$$\left\|g_{n-2j}^{\alpha,\beta,j}\right\|_{L^p_{w(\alpha,\beta)}}=\left\|\sum_{k=0}^{2j} b_{k,j}(\alpha,\beta,n-2j)P_{n-2j+k}^{\alpha,\beta}\right\|_{L^p_{w(\alpha,\beta)}}$$
$$\geq C\left|b_{2j,j}(\alpha,\beta,n-2j)\right|n^{-1/2}\geq Cn^{-1/2},$$

whereas by (3.1) and the assumption on j it follows that

$$\left\|g_{n-2j}^{\alpha,\beta,j}\right\|_{L^p_{w(\alpha,\beta)}}\leq Cn^{-1/2}.$$

b) *Laguerre case*. First of all, we prove (1.13) for $\beta=\alpha$, starting with $p=1$ and $\alpha\geq 0$. Defining a convolution operator as

$$T_n^\alpha f=\Gamma(\alpha+1)\sum_{k=0}^{n} c_{k,n}\hat{f}(k)L_k^\alpha=f*k_n^\alpha\qquad\left(f\in L^1_{w(\alpha)}\right),$$

with the kernel function

$$k_n^\alpha(x)=\Gamma(\alpha+1)\sum_{k=0}^{n} c_{k,n}h_k^\alpha L_k^\alpha(1)L_k^\alpha(x)=\sum_{k=0}^{n} c_{k,n}L_k^\alpha(x)\qquad(x\geq 0),$$

the Laguerre convolution theorem [12] and Theorem 1 yield

$$\left\|\sum_{k=0}^{n} c_{k,n}L_k^\alpha\right\|_{L^1_{w(\alpha)}}=\Gamma(\alpha+1)\left\|T_n^\alpha\right\|_{[L^1_{w(\alpha)}]}\tag{4.1}$$
$$\geq C|c_{n,n}|n^{\alpha+1/2}\qquad(\alpha\geq 0).$$

Though the convolution theorem and therefore (4.1) are only given for $\alpha\geq 0$, one can also deduce this assertion for $-1<\alpha<0$. Indeed, using

$$\int_0^x L_k^\alpha(t)t^\alpha dt=\frac{1}{k+\alpha+1}L_k^{\alpha+1}(x)x^{\alpha+1}\qquad(\alpha>-1)$$

(cf. [6, 10.12.(30)]) and Fubini's theorem, one obtains

$$\text{(4.2)}\qquad \left\| \sum_{k=0}^{n} c_{k,n} \frac{1}{k+\alpha+1} L_k^{\alpha+1}(x) \right\|_{L^1_{w(\alpha+1)}}$$
$$\leq \int_0^\infty \int_0^x \left| \sum_{k=0}^{n} c_{k,n} L_k^\alpha(t) t^\alpha \right| dt e^{-x/2}\, dx$$
$$\leq \int_0^\infty \int_t^\infty e^{-x/2}\, dx \Big| \cdots |dt = 2 \left\| \sum_{k=0}^{n} c_{k,n} L_k^\alpha(t) \right\|_{L^1_{w(\alpha)}}.$$

Since $\alpha+1>0$, (4.1) can be applied to the left-hand side of (4.2) so that

$$\text{(4.3)}\qquad \left\| \sum_{k=0}^{n} c_{k,n} L_k^\alpha \right\|_{L^1_{w(\alpha)}} \geq C|c_{n,n}|(n+\alpha+1)^{-1} n^{\alpha+1+1/2}$$
$$\geq C|c_{n,n}| n^{\alpha+1/2} \qquad (-1<\alpha<0).$$

Let us mention that there also holds a fractional version of (4.2), i.e.,

$$\left\| \sum_{k=0}^{n} c_{k,n} \frac{\Gamma(k+\alpha+1)}{\Gamma(k+\alpha+\beta+1)} L_k^{\alpha+\beta} \right\|_{L^1_{w(\alpha+\beta)}} \leq 2^\beta \left\| \sum_{k=0}^{n} c_{k,n} L_k^\alpha \right\|_{L^1_{w(\alpha)}} \qquad (\alpha>-1, \beta>0).$$

The estimates (4.1), (4.3) remain valid if the parameter α of the Laguerre polynomials is replaced by an arbitrary $\beta>-1$, since, in view of (2.13),

$$\text{(4.4)}\qquad \left\| \sum_{k=0}^{n} c_{k,n} L_k^\beta(x) \right\|_{L^1_{w(\alpha)}} = \left\| \sum_{k=0}^{n} c_{k,n} \sum_{j=0}^{k} A_{k-j}^{\beta-\alpha-1} L_j^\alpha(x) \right\|_{L^1_{w(\alpha)}}$$
$$= \left\| \sum_{j=0}^{n-1} \left[\sum_{k=j}^{n} c_{k,n} A_{k-j}^{\beta-\alpha-1} \right] L_j^\alpha(x) + c_{n,n} L_n^\alpha(x) \right\|_{L^1_{w(\alpha)}}$$
$$\geq C|c_{n,n}| n^{\alpha+1/2} \qquad (\alpha>-1).$$

The assertion (1.13) for $1<p\leq\infty$ can now be deduced from (4.4) by applying Lemma 2 with $\gamma=\frac{\alpha}{p}$ to $p_n=\sum_{k=0}^n c_{k,n} L_k^\beta \in P_n$, i.e.,

$$\left\| \sum_{k=0}^{n} c_{k,n} L_k^\beta \right\|_{L^p_{w(\alpha)}} \geq C n^{(\alpha+1)(1/p-1)} \left\| \sum_{k=0}^{n} c_{k,n} L_k^\beta \right\|_{L^1_{w(\alpha)}}$$
$$\geq C|c_{n,n}| n^{(\alpha+1)/p-1/2}.$$

Concerning (1.14), we replace α by $\alpha p/2$ in (1.13) for $1\leq p<\infty$. For $p=\infty$, we apply Lemma 2 to (4.4) once more, setting $\gamma=\frac{\alpha}{2}$ now.

In order to show that (1.13) is best possible, choose, e.g., the test functions (1.17) with n replaced by $n-2-j$ and α by $\beta>-1$, thus

$$g_{n-2-j}^{\beta,j}(x) = L_{n-2-j}^{\beta+j}(x) x^j - \left(\frac{(n-j-1)(n-j)}{(n+\beta-1)(n+\beta)} \right)^{1/2} L_{n-j}^{\beta+j}(x) x^j,$$

where $n \geq j+3, j > \beta - \frac{1}{2} - (2\alpha+2)/p, j \in \mathbb{P}$. In view of (2.15), it follows by (1.13) that

$$\left\| g_{n-2-j}^{\beta,j} \right\|_{L^p_{w(\alpha)}} = \left\| \sum_{k=0}^{j+2} b_{k,j}(\beta, n-2-j) L_{n-2-j+k}^{\beta} \right\|_{L^p_{w(\alpha)}}$$

$$\geq C \left| b_{j+2,j}(\beta, n-2-j) \right| n^{(\alpha+1)/p - 1/2}$$

$$\geq C n^{j+(\alpha+1)/p-1/2} \qquad (\alpha > -1),$$

and this bound can be seen to be sharp by considering the appropriate modification of (3.3).

Similarly, applying (1.14) to the above test functions with $j > \beta - \alpha - \frac{1}{2} - \frac{2}{p}$, $j \in \mathbb{P}$, one obtains

$$\left\| g_{n-2-j}^{\beta,j} \right\|_{L^p_{u(\alpha)}} \geq C n^{j+1/p+(\alpha-1)/2},$$

the exponent of which is the correct one in view of (1.19) and (2.12).

c) *Hermite case*. Since for each $k \in \mathbb{P}$ [23, (5.6.1)]

$$H_{2k}(x) = (-1)^k 2^{2k} k! L_k^{-1/2}(x^2), \qquad H_{2k+1}(x) = (-1)^k 2^{2k+1} k! L_k^{1/2}(x^2) x,$$

the polynomial $p_n(x) = \sum_{k=0}^{n} c_{k,n} H_k(x)$ can be divided into its even and odd parts by

$$p_n(x) = \sum_{k=0}^{[n/2]} c_{2k,n} (-1)^k 2^{2k} k! L_k^{-1/2}(x^2)$$

$$+ \sum_{k=0}^{[(n-1)/2]} c_{2k+1,n} (-1)^k 2^{2k+1} k! L_k^{1/2}(x^2) x$$

$$= p_n^e(x) + p_n^o(x),$$

say. Taking norms, we have

$$\|p_n\|_{L^1_{u(H)}} = \int_{-\infty}^{\infty} |p_n(x)| e^{-x^2/2} dx \tag{4.5}$$

$$= \int_0^{\infty} |p_n^e(x) + p_n^o(x)| e^{-x^2/2} dx + \int_0^{\infty} |p_n^e(x) - p_n^o(x)| e^{-x^2/2} dx$$

$$\geq 2 \max \left\{ \int_0^{\infty} |p_n^e(x)| e^{-x^2/2} dx, \int_0^{\infty} |p_n^o(x)| e^{-x^2/2} dx \right\}.$$

If $n = 2m$, $m \in \mathbb{P}$, we use the even part estimate of (4.5), set $x^2 = y$ and apply (1.13) for $\alpha = \beta = -\frac{1}{2}$ to derive

$$\|p_n\|_{L^1_{u(H)}} \geq 2 \int_0^{\infty} |p_n^e(x)| e^{-x^2/2} dx$$

$$= \left\| \sum_{k=0}^{m} c_{2k,n} (-1)^k 2^{2k} k! L_k^{-1/2}(y) \right\|_{L^1_{w(-1/2)}} > C |c_{n,n}| 2^n m!$$

If $n=2m+1$, $m\in\mathbb{P}$, we use the odd part estimate of (4.5), set $x^2=y$ again, and apply (1.13) for $\beta=\frac{1}{2}$, $\alpha=0$ to deduce

$$\begin{aligned}\|p_n\|_{L^1_{u(H)}} &\geq 2\int_0^\infty |p_n^o(x)|e^{-x^2/2}dx \\ &= \left\| \sum_{k=0}^{m} c_{2k+1,n}(-1)^k 2^{2k+1}k!L_k^{1/2}(y)\right\|_{L^1_{w(0)}} \\ &\geq C|c_{n,n}|2^n m!m^{1/2}.\end{aligned}$$

In view of Legendre's duplication formula for the gamma function, $\Gamma(2z)=2^{2z-1}\pi^{-1/2}\Gamma(z)\Gamma(z+\frac{1}{2})$, one has

$$\begin{aligned}\left(\sqrt{\pi}\,2^n n!\right)^{1/2} &= \left(2^{2n}\Gamma((n+1)/2)\Gamma(n/2+1)\right)^{1/2} \\ &= 2^n\begin{cases} m!\left(\Gamma(m+1/2)/\Gamma(m+1)\right)^{1/2}, & n=2m, \\ m!\left(\Gamma(m+3/2)/\Gamma(m+1)\right)^{1/2}, & n=2m+1 \end{cases} \qquad (m\in\mathbb{P}).\end{aligned}$$

Thus, for each n, one obtains

$$\left\| \sum_{k=0}^{n} c_{k,n}H_k\right\|_{L^1_{u(H)}} \geq C|c_{n,n}|\left(\sqrt{\pi}\,2^n n!\right)^{1/2} n^{1/4}. \tag{4.6}$$

Applying the Nikolskii type inequality

$$\|p_n\|_{L^1_{u(H)}} \leq Cn^{(1/2)(1-1/p)}\|p_n\|_{L^p_{u(H)}} \qquad (p_n\in P_n, 1\leq p\leq\infty),$$

(cf. [8]) to the left-hand side of (4.6), we arrive at (1.15).

These estimates are best possible since, when applied to the test functions (1.20) with n replaced by $n-2$, they yield

$$\begin{aligned}\left\|g_{n-2}^H(x)\right\|_{L^p_{u(H)}} &\geq C[4(n-1)n]^{-1/2}\left(\sqrt{\pi}\,2^n n!\right)^{1/2} n^{1/(2p)-1/4} \\ &= C\left(\sqrt{\pi}\,2^{n-2}(n-2)!\right)^{1/2} n^{1/(2p)-1/4}.\end{aligned}$$

On the other hand one has, in view of (2.17),

$$\begin{aligned}\left\|g_{n-2}^H(x)\right\|_{L^p_{u(H)}} &= \left(\sqrt{\pi}\,2^{n-2}(n-2)!\right)^{1/2}\left\|\mathcal{H}_{n-2}(x)-\mathcal{H}_n(x)\right\|_{L^p(-\infty,\infty)} \\ &\leq C\left(\sqrt{\pi}\,2^{n-2}(n-2)!\right)^{1/2} n^{1/(2p)-1/4}.\end{aligned}$$

□

Acknowledgment. The author would like to thank Professor E. Görlich for many helpful comments and suggestions.

REFERENCES

[1] R. Askey, *Dual equations and classical orthogonal polynomials*, J. Math. Anal. Appl., 24 (1968), pp. 677–685.

[2] R. Askey and S. Wainger, *Mean convergence of expansions in Laguerre and Hermite series*, Amer. J. Math., 87 (1965), pp. 695–708.

[3] P. J. Cohen, *On a conjecture of Littlewood and idempotent measures*, Amer. J. Math., 82 (1960), pp. 191–212.

[4] B. DRESELER AND P. M. SOARDI, *A Cohen type inequality for ultraspherical series*, Arch. Math., 38 (1982), pp. 243–247.

[5] _____, *A Cohen type inequality for Jacobi expansions and divergence of Fourier series on compact symmetric spaces*, J. Approximation Theory, 35 (1982), pp. 214–221.

[6] A. ERDÉLYI ET AL., *Higher Transcendental Functions*, Vols. I, II, McGraw-Hill, New York, 1953.

[7] J. J. F. FOURNIER, *On a theorem of Paley and the Littlewood conjecture*, Ark. Mat., 17 (1979), pp. 199–216.

[8] G. GASPER, *Positivity and the convolution structure for Jacobi series*, Ann. Math., 93 (1971), pp. 112–118.

[9] S. GIULINI, P. M. SOARDI, AND G. TRAVAGLINI, *A Cohen type inequality for compact Lie groups*, Proc. Amer. Math. Soc., 77 (1979), pp. 359–364.

[10] E. GÖRLICH AND C. MARKETT, *On a relation between the norms of Cesàro means of Jacobi expansions*, Linear Spaces and Approximation, P. L. Butzer, B. Sz.-Nagy, eds., ISNM 40, Birkhäuser Verlag, Basel, 1978, pp. 251–262.

[11] _____, *Projections with norms smaller than those of the ultrasperical and Laguerre partial sums*, Functional Analysis and Approximation, P. L. Butzer, B. Sz.-Nagy, E. Görlich, eds., ISNM 60, Birkhäuser Verlag, Basel, 1981, pp. 189–202.

[12] _____, *A convolution structure for Laguerre series*, Indag. Math., ser. A, 85 (1982), pp. 161–171.

[13] G. H. HARDY AND J. E. LITTLEWOOD, *A new proof of a theorem on rearrangements*, J. London Math. Soc., 23 (1948), pp. 163–168.

[14] S. G. KAL'NEI, *Uniform boundedness in the L-metric of polynomials with respect to the Jacobi polynomials*, Dokl. Akad. Nauk SSSR 222 (1975), = Soviet Math. Dokl., 16 (1975), pp. 714–718.

[15] C. MARKETT, *Norm estimates for Cesàro means of Laguerre expansions*, Approximation and Function Spaces, Z. Ciesielski, ed., North-Holland, Amsterdam, 1981, pp. 419–435.

[16] _____, *Mean Cesàro summability of Laguerre expansions and norm estimates with shifted parameter*, Anal. Math., 8 (1982), pp. 19–37.

[17] _____, *Norm estimates for (C,δ) means of Hermite expansions and bounds for δ_{eff}*, to appear.

[18] _____, *Nikolskii type inequalities for Laguerre and Hermite expansions*, Colloq. Math. Soc. János Bolyai 35, Functions, Series, Operators, J. Szabados, ed., North-Holland, Amsterdam, 1982.

[19] O. C. MCGEHEE, L. PIGNO, AND B. SMITH, *Hardy's inequality and the Littlewood conjecture*, Ann. Math., to appear. An announcement of the results appeared in Bull. Amer. Math. Soc., 5 (1981), pp. 71–72.

[20] CH. MEANEY, *Unbounded Lebesgue constants on compact groups*, Mh. Math. 91 (1981), pp. 119–129.

[21] B. MUCKENHOUPT, *Mean convergence of Jacobi series*. Proc. Amer. Math. Soc., 23 (1969), pp. 306–310.

[22] _____, *Mean convergence of Hermite and Laguerre series* II., Trans. Amer. Math. Soc., 147 (1970), pp. 433–460.

[23] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, R.I., 1975.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1404-0016 $01.25/0

WEIGHTED NORM INEQUALITIES FOR CERTAIN INTEGRAL OPERATORS*

K. F. ANDERSEN† AND H. P. HEINIG‡

Abstract. Conditions on the nonnegative weight functions $u(x)$ and $v(x)$ are given which ensure that an inequality of the form $(\int|(Tf)(x)u(x)|^q\,dx)^{1/q}\le C(\int|f(x)v(x)|^p\,dx)^{1/p}$ holds where T is an integral operator of the form $\int_{-\infty}^{x}K(x,y)f(y)\,dy$ or $\int_x^{\infty}K(x,y)f(y)\,dy$ and C is a constant depending on K,p,q but independent of f; the inequality being reversed in case $p,q<1$. In particular, new inequalities and a unified treatment of several known inequalities are obtained for a class of convolution operators, various fractional integrals and the Laplace transform.

1. Introduction. Let $K(x,y)\ge 0$ be defined on $\Delta=\{(x,y)\in R^2: y<x\}$ and define the operator K and its dual K^* by

$$(1.1)\qquad (Kf)(x)=\int_{-\infty}^{x}K(x,y)f(y)\,dy,\qquad (K^*f)(x)=\int_x^{\infty}K(y,x)f(y)\,dy.$$

The purpose of this paper is to give conditions on the nonnegative weight functions $u(x)$, $v(x)$ in terms of the kernel $K(x,y)$ and the indices p,q which imply inequalities of the form

$$\left(\int|(Tf)(x)u(x)|^q dx\right)^{1/q}\le C\left(\int|f(x)v(x)|^p dx\right)^{1/p},$$

where T is either K or K^* and C is a constant independent of f; the sense of the inequality being reversed if $p,q<1$.

The particular case of the Hardy operators given by the kernel $K(x,y)\equiv 1$ was considered by Andersen and Muckenhoupt [1] and independently by Bradley [3] for the case $p,q\ge 1$, while Beesack and Heinig [2] dealt with the case $p,q<1$. This paper extends these results to a general class of kernels $K(x,y)$.

The class of operators considered here includes several classical operators, among them the Laplace transform and the fractional integrals of Riemann–Liouville and Weyl. For these, our results at once provide new inequalities and a unified treatment of various known inequalities, particularly those involving power weights as given in [4].

Applications to other operators, including a class of convolution integrals, are also given. The discrete analogues of our integral inequalities are briefly considered; these generalize results of Leindler [5].

The plan of the paper is as follows: In the next section we prove our principal results (Theorems 2.1 and 2.2) for the case $p,q\ge 1$ and give various applications. Section 3 contains the integral estimates for the cases $0<p,q<1$ and $p,q<0$. In the last section we discuss the discrete cases.

Throughout, p' denotes the conjugate index of $p, p\ne 0$ and is defined by $1/p+1/p'=1$ with $p'=\infty$ if $p=1$. The conjugate of q is defined in the same way. Furthermore, products of the form $0\cdot\infty$ are taken to be zero. A,B and C denote constants

* Received by the editors June 22, 1981, and in revised form June 1, 1982.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The research of this author was supported in part by the Natural Sciences and Engineering Research Council grant A-8185.

‡ Department of Mathematical Sciences, McMaster University, Hamilton, Ontario, Canada L8S 4K1. The research of this author was supported in part by the Natural Sciences and Engineering Research Council grant A-4837.

which may be different at different occurrences, while Z and Z^+ denote the integers and the positive integers respectively.

2. Weighted integral inequalities for $1\le p\le q\le\infty$. The weight functions we consider depend on the kernel of the integral operator.

DEFINITION 2.1. Let $u(x)\ge 0$, $v(x)\ge 0$, $1\le p\le q\le\infty$, and suppose $K(x,y)\ge 0$ is defined in $\Delta=\{(x,y)\in R^2: y<x\}$.

(a) We say (u,v) satisfies $A(K,p,q)$ with constant C if there is a β, $0\le\beta\le 1$, such that for all real a

$$\left(\int_a^\infty \left[K(y,a)^\beta u(y)\right]^q dy\right)^{1/q}\left(\int_{-\infty}^a \left[K(a,y)^{\beta-1}v(y)\right]^{-p'}dy\right)^{1/p'}\le C<\infty, \tag{2.1}$$

holds. If $1=p\le q<\infty$ (2.1) takes the form

$$\left(\int_a^\infty \left[K(y,a)^\beta u(y)\right]^q dy\right)^{1/q}\operatorname*{ess\,sup}_{y<a}\left[K(a,y)^{\beta-1}v(y)\right]^{-1}\le C, \tag{2.2}$$

with the usual modification if $q=\infty$.

(b) We say (u,v) satisfies $A^*(K,p,q)$ with constant C^* if there is a β, $0\le\beta\le 1$, such that for all real a

$$\left(\int_{-\infty}^a \left[K(a,y)^\beta u(y)\right]^q dy\right)^{1/q}\left(\int_a^\infty \left[K(y,a)^{\beta-1}v(y)\right]^{-p'}dy\right)^{1/p'}\le C^*<\infty \tag{2.3}$$

holds. For $1=p\le q<\infty$ the modification in (2.3) is similar to that of (2.2), except that the essential supremum is over $y>a$.

Before stating the main result of this section we give the following integral form of Minkowski's inequality [4, Thm. 202, p. 148].

LEMMA 2.1. *If $g(x,y)\ge 0$, $1\le p\le\infty$, and $b\ge -\infty$, then*

$$\left(\int_b^\infty\left[\int_b^x g(x,y)\,dy\right]^p dx\right)^{1/p}\le\int_b^\infty\left[\int_y^\infty g(x,y)^p dx\right]^{1/p}dy \tag{2.4}$$

and

$$\left(\int_b^\infty\left[\int_x^\infty g(x,y)\,dy\right]^p dx\right)^{1/p}\le\int_b^\infty\left[\int_b^y g(x,y)^p dx\right]^{1/p}dy, \tag{2.5}$$

with the usual modification if $p=\infty$. If $p<1$ the inequalities in (2.4) *and* (2.5) *are reversed.*

THEOREM 2.1. *Let K be the integral operator defined by* (1.1) *where $K(x,y)\ge 0$ is defined in Δ with $K(x,y)$ nonincreasing in x and nondecreasing in y. If $1\le p\le q<\infty$ and (u,v) satisfies $A(K,p,q)$ with constant C, then*

$$\left(\int_{-\infty}^\infty |u(x)(Kf)(x)|^q dx\right)^{1/q}\le AC\left(\int_{-\infty}^\infty |v(x)f(x)|^p dx\right)^{1/p}, \tag{2.6}$$

where $A=((p'+q)/q)^{1/p'}((p'+q)/p')^{1/q}$ if $1<p\le q<\infty$ and $A=1$ otherwise.

Proof. Assume $f\ge 0$ for which the right side of (2.6) is finite.

Consider first the case $1<p\le q<\infty$, and define h by

$$h(y)=\left(\int_{-\infty}^y K(y,z)^{(1-\beta)p'}v(z)^{-p'}dz\right)^{1/(p'+q)}.$$

If the left side of (2.6) is denoted by I, Hölder's inequality shows that

$$I\le\left(\int_{-\infty}^{\infty}u(x)^q\left(\int_{-\infty}^{x}\left[K(x,y)^{\beta}f(y)v(y)h(y)\right]^p dy\right)^{q/p}\right.$$
$$\left.\cdot\left(\int_{-\infty}^{x}\left[K(x,y)^{\beta-1}v(y)h(y)\right]^{-p'}dy\right)^{q/p'}dx\right)^{1/q}.$$

The second inner integral has the form

$$\int_{-\infty}^{x}K(x,y)^{(1-\beta)p'}v(y)^{-p'}\left(\int_{-\infty}^{y}K(y,z)^{(1-\beta)p'}v(z)^{-p'}dz\right)^{-p'/(p'+q)}dy,$$

and since $y<x$ implies $K(y,z)\ge K(x,z)$, this is bounded above by

$$\int_{-\infty}^{x}K(x,y)^{(1-\beta)p'}v(y)^{-p'}\left(\int_{-\infty}^{y}K(x,z)^{(1-\beta)p'}v(z)^{-p'}dz\right)^{-p'/(p'+q)}dy \tag{2.7}$$
$$=\left(\frac{p'+q}{q}\right)\left(\int_{-\infty}^{x}K(x,z)^{(1-\beta)p'}v(z)^{-p'}dz\right)^{q/(p'+q)}=\left(\frac{p'+q}{q}\right)h(x)^q,$$

which was obtained by integrating. Hence (2.4) and the definition of $A(K,p,q)$ yield

$$I\le\left(\frac{p'+q}{q}\right)^{1/p'}\left(\int_{-\infty}^{\infty}\left[u(x)h(x)^{q/p'}\right]^q\left(\int_{-\infty}^{x}K(x,y)^{\beta p}\left[f(y)v(y)h(y)\right]^p dy\right)^{q/p}dx\right)^{1/q}$$
$$\le\left(\frac{p'+q}{q}\right)^{1/p'}\left(\int_{-\infty}^{\infty}\left[f(y)v(y)h(y)\right]^p\left(\int_{y}^{\infty}K(x,y)^{\beta q}u(x)^q h(x)^{q^2/p'}dx\right)^{p/q}dy\right)^{1/p}$$
$$\le\left(\frac{p'+q}{q}\right)^{1/p'}C^{q/(p'+q)}$$
$$\cdot\left(\int_{-\infty}^{\infty}\left[f(y)v(y)h(y)\right]^p\left(\int_{y}^{\infty}K(x,y)^{\beta q}u(x)^q\right.\right.$$
$$\left.\left.\cdot\left(\int_{x}^{\infty}K(z,x)^{\beta q}u(z)^q dz\right)^{-q/(p'+q)}dx\right)^{p/q}dy\right)^{1/p}.$$

Again $y<x$ implies $K(z,y)\le K(z,x)$, so that on replacing $K(z,x)$ by $K(z,y)$ in the inner integral and then integrating, one obtains

$$I\le\left(\frac{p'+q}{q}\right)^{1/p'}C^{q/(p'+q)}\left(\frac{p'+q}{p'}\right)^{1/q}\left(\int_{-\infty}^{\infty}\left[f(y)v(y)h(y)\right]^p\right.$$
$$\left.\cdot\left(\int_{y}^{\infty}K(z,y)^{\beta q}u(z)^q dz\right)^{pp'/q(p'+q)}dy\right)^{1/p}$$
$$\le\left(\frac{p'+q}{q}\right)^{1/p'}\left(\frac{p'+q}{p'}\right)^{1/q}C\left(\int_{-\infty}^{\infty}\left[f(y)v(y)\right]^p dy\right)^{1/p},$$

where we used (2.1) and the definition of h. This completes the proof if $1<p\le q<\infty$.

Now suppose $1<p\le q=\infty$. For each x we have

$$u(x)\int_{-\infty}^{x}K(x,y)f(y)\,dy\le u(x)\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}\int_{-\infty}^{x}K(x,y)^{1-\beta}f(y)\,dy,$$

so that Hölder's inequality and the definition of $A(K,p,\infty)$ show that this is bounded above by

$$u(x)\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}\left(\int_{-\infty}^{x}\left[K(x,y)^{\beta-1}v(y)\right]^{-p'}dy\right)^{1/p'}\left(\int_{-\infty}^{x}\left[f(y)v(y)\right]^{p}dy\right)^{1/p}$$

$$\leq C\left(\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}u(x)\right)\left(\operatorname*{ess\,sup}_{t>x}K(t,x)^{\beta}u(t)\right)^{-1}\left(\int_{-\infty}^{\infty}\left[f(y)v(y)\right]^{p}dy\right)^{1/p},$$

so that it suffices to prove that the product of the first two factors on the right side is bounded above by 1. To see this, observe that

$$\left(\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}u(x)\right)\left(\operatorname*{ess\,sup}_{t>x}K(t,x)^{\beta}u(t)\right)^{-1}$$

$$=\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}u(x)\left(\operatorname*{ess\,sup}_{t>x}K(t,x)^{\beta}u(t)\right)^{-1}$$

$$\leq\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}u(x)\left(\operatorname*{ess\,sup}_{t>x}K(t,z)^{\beta}u(t)\right)^{-1}$$

since $K(t,z)\leq K(t,x)$ for $z<x$. This is now clearly bounded above by

$$\operatorname*{ess\,sup}_{z<x}K(x,z)^{\beta}u(x)\left(K(x,z)^{\beta}u(x)\right)^{-1}=1,$$

as required. This completes the proof if $1<p\leq q=\infty$.

If $1=p\leq q$, define h by

$$h(y)=\operatorname*{ess\,sup}_{z<y}\left[K(y,z)^{\beta-1}v(z)\right]^{-1}.$$

Again (since $x>y$), the monotonicity condition on K implies that

$$I=\left(\int_{-\infty}^{\infty}u(x)^{q}\left(\int_{-\infty}^{x}K(x,y)f(y)h(y)\operatorname*{ess\,inf}_{z<y}\left[K(y,z)^{\beta-1}v(z)\right]dy\right)^{q}dx\right)^{1/q}$$

$$\leq\left(\int_{-\infty}^{\infty}u(x)^{q}\left(\int_{-\infty}^{x}K(x,y)f(y)h(y)\operatorname*{ess\,inf}_{z<y}\left[K(x,z)^{\beta-1}v(z)\right]dy\right)^{q}dx\right)^{1/q}$$

$$\leq\left(\int_{-\infty}^{\infty}u(x)^{q}\left(\int_{-\infty}^{x}K(x,y)^{\beta}f(y)v(y)h(y)\,dy\right)^{q}dx\right)^{1/q}.$$

We now apply (2.4) and (2.2) to obtain

$$I\leq\int_{-\infty}^{\infty}f(y)v(y)h(y)\left(\int_{y}^{\infty}K(x,y)^{\beta q}u(x)^{q}dx\right)^{1/q}dy$$

$$\leq C\int_{-\infty}^{\infty}f(y)v(y)\,dy.$$

It is clear that this last argument also holds if $q=\infty$. This completes the proof of Theorem 2.1.

Our next result is the dual of Theorem 2.1.

THEOREM 2.2. *Let K^* be the integral operator defined by* (1.1) *where $K(x,y)\geq 0$ is defined in Δ with $K(x,y)$ nonincreasing in x and nondecreasing in y. If $1\leq p\leq q\leq\infty$ and*

(u,v) satisfies $A^*(K,p,q)$ with constant C^*, then

$$\left(\int_{-\infty}^{\infty}|u(x)(K^*f)(x)|^q dx\right)^{1/q}\le AC^*\left(\int_{-\infty}^{\infty}|v(x)f(x)|^p dx\right)^{1/p} \tag{2.8}$$

where $A=((p'+q)/q)^{1/p'}((p'+q)/p')^{1/q}$ if $1<p\le q<\infty$ and $A=1$ otherwise.

Proof. Hölder's inequality and its converse show that (2.8) is equivalent to

(2.9)

$$\int_{-\infty}^{\infty}g(x)(K^*f)(x)\,dx\le AC^*\left(\int_{-\infty}^{\infty}[v(x)f(x)]^p dx\right)^{1/p}\left(\int_{-\infty}^{\infty}g(x)^{q'}u(x)^{-q'}dx\right)^{1/q'}$$

for all nonnegative f and g. Fubini's theorem shows that the left side equals $\int_{-\infty}^{\infty}f(y)(Kg)(y)\,dy$ so that Theorem 2.1 implies that (2.9) holds if (v^{-1},u^{-1}) satisfies $A(K,q',p')$ with constant C^*; that is, if (u,v) satisfies $A^*(K,p,q)$ with constant C^*. This proves Theorem 2.2.

An important class of operators are those given by convolution integrals of the form

$$\int_{-\infty}^{x}k(x-y)f(y)\,dy \quad\text{and}\quad \int_{x}^{\infty}k(y-x)f(y)\,dy.$$

For these we have the following result.

COROLLARY 2.1. *Let $1\le p\le q\le\infty$, $k(x)\ge 0$, be nonincreasing, and suppose that $u\ge 0$, $v\ge 0$ satisfy*

$$\sup_r\left(\int_r^{\infty}k(x-r)^{\beta q}u(x)^q dx\right)^{1/q}\left(\int_{-\infty}^{r}k(r-x)^{(1-\beta)p'}v(x)^{-p'}dx\right)^{1/p'}\equiv C<\infty \tag{2.10}$$

for some $0\le\beta\le 1$. Then

$$\left(\int_{-\infty}^{\infty}\left|u(x)\int_{-\infty}^{x}k(x-y)f(y)\,dy\right|^q dx\right)^{1/q}\le AC\left(\int_{-\infty}^{\infty}|f(x)v(x)|^p dx\right)^{1/p}, \tag{2.11}$$

where $A=((p'+q)/q)^{1/p'}((p'+q)/p')^{1/q}$ if $1<p\le q<\infty$ and $A=1$ otherwise.

There is a similar result for the dual operator.

The corollary and its dual follow at once from Theorems 2.1 and 2.2.

If we take $k\equiv 1$ and f supported on $(0,\infty)$, then we obtain the generalizations of Hardy's inequality given by Andersen and Muckenhoupt [1] and Bradley [3]. In fact, in this case (2.11) implies (2.10) and similarly for the dual.

From Corollary 2.1 and its dual one easily obtains inequalities for fractional integral operators and some of their generalizations, such as the Erdélyi–Kober operators $I^{\nu}_{\alpha,\xi}$ and $J^{\nu}_{\beta,\eta}$ defined by

$$(I^{\nu}_{\alpha,\xi}f)(x)=\frac{\nu x^{-\nu(\xi+\alpha)+\nu}}{\Gamma(\alpha)}\int_0^x(x^{\nu}-y^{\nu})^{\alpha-1}y^{\nu\xi-1}f(y)\,dy,$$

$$(J^{\nu}_{\beta,\eta}f)(x)=\frac{\nu x^{\nu\eta}}{\Gamma(\alpha)}\int_x^{\infty}(y^{\nu}-x^{\nu})^{\beta-1}y^{-\nu(\beta+\eta)+\nu-1}f(y)\,dy.$$

We give the details only for $I^{\nu}_{\alpha,\xi}$ since $J^{\nu}_{\beta,\eta}$ may be treated by a dual argument. Elementary variable changes show that the inequality

$$\left(\int_0^{\infty}|u(x)(I^{\nu}_{\alpha,\xi}f)(x)|^q dx\right)^{1/q}\le C\left(\int_0^{\infty}|f(x)v(x)|^p dx\right)^{1/p}$$

is equivalent to

$$\left(\int_0^\infty |U(x)(I_\alpha f)(x)|^q dx\right)^{1/q} \le C\left(\int_0^\infty |f(x)V(x)|^p dx\right)^{1/p}$$

where $U(x)=x^{1-\xi-\alpha+(1/\nu-1)/q}u(x^{1/\nu})$, $V(x)=x^{1-\xi+(1/\nu-1)/p}v(x^{1/\nu})$ and I_α denotes the Riemann–Liouville fractional integral $(I_\alpha f)(x)=x^\alpha(I^1_{\alpha,1}f)(x)$. For I_α, Corollary 2.1 yields the following result:

THEOREM 2.3. *Suppose* $1\le p\le q\le\infty$, $0<\alpha<1$, *and there is* β, $0\le\beta\le 1$, *satisfying*

$$\left(\int_r^\infty \left[(x-r)^{(\alpha-1)\beta}U(x)\right]^q dx\right)^{1/q}\left(\int_0^r (r-x)^{(\alpha-1)(1-\beta)p'}V(x)^{-p'}dx\right)^{1/p'} \le C \tag{2.12}$$

for all $r>0$. *Then there is a constant* $B>0$ *such that*

$$\left(\int_0^\infty |U(x)(I_\alpha f)(x)|^q dx\right)^{1/q} \le B\left(\int_0^\infty |f(x)V(x)|^p dx\right)^{1/p}. \tag{2.13}$$

Theorem 2.3 may be applied, in particular, to obtain weight functions $U(x)$, $V(x)$, each of the form $x^a(1+x)^{-a-b}$, for which (2.13) holds. For simplicity, we illustrate this only for the case of power weights.

COROLLARY 2.2. *Suppose* $1<p\le q<\infty$, $1/q=1/p+\gamma+\delta-\alpha$, $\gamma>0$, $\delta<1/p'$ *and* $0<\gamma+\delta\le\alpha\le 1$. *There is a constant* $B>0$ *such that*

$$\left(\int_0^\infty |x^{-\gamma}(I_\alpha f)(x)|^q dx\right)^{1/q} \le B\left(\int_0^\infty |f(x)x^\delta|^p dx\right)^{1/p}.$$

If we take $1<p=q$, $\delta=0$, $0<\alpha<1$, we obtain [4, Thm. 329]; the choice $1<p$, $0<\alpha<1/p$, $\gamma=(p-q+pq\alpha)/(pq)$, $\gamma>0$, $\delta=0$ yields [4, Thm. 402] for the range $p\le q<p/(1-\alpha p)$.

Proof of Corollary 2.2. The result follows from Theorem 2.3 if we show that for all $r>0$

$$\left(\int_r^\infty (x-r)^{(\alpha-1)\beta q}x^{-\gamma q}dx\right)^{1/q}\left(\int_0^r (r-x)^{(1-\beta)(\alpha-1)p'}x^{-\delta p'}dx\right)^{1/p'} \le C<\infty \tag{2.14}$$

for a suitable choice of β, $0\le\beta\le 1$, to be selected presently. Let $x=r/t$ in the first integral and $x=rt$ in the second. Then the left side of (2.14) becomes $r^{-\gamma+1/q+\alpha-1-\delta+1/p'}$ multiplied by the product

$$\left(\int_0^1 (1-t)^{(\alpha-1)\beta q}t^{(1-\alpha)\beta q+\gamma q-2}dt\right)^{1/q}\left(\int_0^1 (1-t)^{(1-\beta)(\alpha-1)p'}t^{-\delta p'}dt\right)^{1/p'}.$$

Now the definition of q shows that the exponent of r is zero, while the integrals converge if $(\alpha-1)\beta q+1>0$, $(1-\alpha)\beta q+\gamma q-1>0$, $(1-\beta)(\alpha-1)p'+1>0$ and $\delta p'<1$. Since $p>1$ we have $1/p-\alpha<1-\alpha$, and since $\delta<1/p'$ it follows that $1/q-\gamma=1/p+\delta-\alpha=\delta-1/p'+1-\alpha<1-\alpha$, while $\gamma>0$ and $\gamma+\delta>0$ imply that $1/q>\max(1/q-\gamma, 1/p-\alpha)$. Hence, there is a β, $0\le\beta\le 1$, satisfying $\max(1/q-\gamma, 1/p-\alpha)<(1-\alpha)\beta<1/q$. For this β the integrals converge and the corollary follows.

Observe that

$$\int_0^x \frac{|f(y)|}{(x-y)^{1-\alpha}}dy \ge h^{\alpha-1}\int_{x-h}^x |f(y)|\,dy \qquad (0<h<x)$$

so that the left fractional maximal function $M_\alpha f$ given by

$$(M_\alpha f)(x)=\sup_{0<h<x} h^{\alpha-1}\int_{x-h}^{x}|f(y)|\,dy$$

satisfies $M_\alpha f\le\Gamma(\alpha)(I_\alpha|f|)$. As a consequence, Theorem 2.3 yields two weight function inequalities for M_α also. A dual argument yields similar inequalities for the right fractional maximal function.

As a final application of Theorems 2.1 and 2.2, consider the Laplace transform $\mathfrak{L}$ given by

$$(\mathfrak{L}f)(x)=\int_0^\infty e^{-xy}f(y)\,dy \qquad (x>0).$$

If

$$(Kf)(x)=\int_0^x f(y)\,dy \quad\text{and}\quad (K_1^*f)(x)=\int_x^\infty e^{-y/x}f(y)\,dy,$$

then for $f(y)\ge 0$

$$e^{-1}(Kf)\left(\frac{1}{x}\right)+(K_1^*f)\left(\frac{1}{x}\right)\le(\mathfrak{L}f)(x)\le(Kf)\left(\frac{1}{x}\right)+(K_1^*f)\left(\frac{1}{x}\right).$$

It follows that

$$\left(\int_0^\infty|u(x)(\mathfrak{L}f)(x)|^q dx\right)^{1/q}\le C\left(\int_0^\infty|f(x)v(x)|^p dx\right)^{1/p}$$

if and only if both

$$\int_0^\infty\left|u\left(\frac{1}{x}\right)(Kf)(x)\right|^q\frac{dx}{x^2} \quad\text{and}\quad \int_0^\infty\left|u\left(\frac{1}{x}\right)(K_1^*f)(x)\right|^q\frac{dx}{x^2}$$

are bounded by a multiple of

$$\left(\int_0^\infty|f(x)v(x)|^p dx\right)^{q/p}.$$

Thus, Theorems 2.1 and 2.2 yield the following result.

THEOREM 2.4. *Suppose* $1\le p\le q\le\infty$ *and for all* $r>0$

$$\left(\int_0^{1/r}u(x)^q dx\right)^{1/q}\left(\int_0^r v(x)^{-p'}dx\right)^{1/p'}\le C. \tag{2.15}$$

If for some β, $0\le\beta\le 1$,

$$\left(\int_{1/r}^\infty e^{-\beta rqx}u(x)^q dx\right)^{1/q}\left(\int_r^\infty e^{-(1-\beta)p'x/r}v(x)^{-p'}dx\right)^{1/p'}\le C \tag{2.16}$$

for all $r>0$, *then*

$$\left(\int_0^\infty|u(x)(\mathfrak{L}f)(x)|^q dx\right)^{1/q}\le C\left(\int_0^\infty|f(x)v(x)|^p dx\right)^{1/p}. \tag{2.17}$$

The conditions (2.15) and (2.16) imply the required inequalities for K and K_1^* respectively. Moreover, as we observed following Corollary 2.1, (2.15) is also a necessary condition for the inequality involving K, and hence is also a necessary condition for (2.17).

If $u(x)$ is nonincreasing and $v(x)$ is nondecreasing, then (2.15) is both necessary and sufficient for (2.17). To see this, observe that (2.15) and the monotone conditions

imply that $[u(1/r)^q/r]^{1/q}[v(r)^{-p'}r]^{1/p'} \le C$, while for $0<\beta<1$ the left side of (2.16) is clearly bounded above by

$$\left[e^{-\beta q}u(1/r)^q/\beta qr\right]^{1/q}\left[e^{-(1-\beta)p'}v(r)^{-p'}r/(1-\beta)p'\right]^{1/p'}.$$

Thus (2.15) implies (2.16) in this case, and the result follows.

It is easy to verify that $u(x)=x^{a-1/q}$, $v(x)=x^{-a+1/p'}$ satisfy the requirements of Theorem 2.4 provided $a>0$. The particular case $a=1/p'$, $1<p\le q<\infty$, is [4, Thm. 360]. The case $1<p=q<\infty$, $a=1/p$ or $1/p'$, is [4, Thm. 350] while the case $1<p\le 2$, $q=p'$, $a=1/p$ is given as [4, Thm. 352].

3. The integral inequalities for $p,q<1$. In this section we prove results corresponding to those of the previous section for indices less than 1.

Let $u(x)\ge 0$, $v(x)\ge 0$, $K(x,y)\ge 0$, $p,q<1$, $p\ne 0$, $q\ne 0$ and $\beta=0$ or 1. For any real r we define K_β and J_β by

$$K_\beta(r)=\left(\int_{-\infty}^{r}\left[K(r,s)^\beta u(s)\right]^q ds\right)^{1/q}\left(\int_{-\infty}^{r}\left[K(r,s)^{\beta-1}v(s)\right]^{-p'}ds\right)^{1/p'}$$

and

$$J_\beta(r)=\left(\int_{r}^{\infty}\left[K(s,r)^\beta u(s)\right]^q ds\right)^{1/q}\left(\int_{r}^{\infty}\left[K(s,r)^{\beta-1}v(s)\right]^{-p'}ds\right)^{1/p'}.$$

DEFINITION 3.1. We say a nonnegative function f is *K-admissible*, respectively *K*-admissible*, if $(Kf)(x)$, respectively $(K^*f)(x)$, is finite for all x.

Our first result deals with the case $0<q\le p<1$.

THEOREM 3.1. *Let $K(x,y)\ge 0$ be defined in Δ and suppose $0<q\le p<1$.*

(a) *If $K(x,y)$ is nondecreasing in y and (u,v) satisfy $\inf_r J_1(r)\equiv B>0$, with $J_1(r)$ either bounded above or nonincreasing, then*

$$\left(\int_{-\infty}^{\infty}\left[v(x)f(x)\right]^p dx\right)^{1/p}\le C\left(\int_{-\infty}^{\infty}\left[u(x)(Kf)(x)\right]^q dx\right)^{1/q} \tag{3.1}$$

holds for all K-admissible f and some constant C.

(b) *If $K(x,y)$ is nonincreasing in x and (u,v) satisfy $\inf_r K_1(r)\equiv B>0$, with $K_1(r)$ either bounded above or nondecreasing, then*

$$\left(\int_{-\infty}^{\infty}\left[v(x)f(x)\right]^p dx\right)^{1/p}\le C\left(\int_{-\infty}^{\infty}\left[u(x)(K^*f)(x)\right]^q dx\right)^{1/q} \tag{3.2}$$

holds for all K-admissible f and some constant C.*

Proof. Assume the right side of (3.1) is finite and f is K-admissible. We shall carry out the proof assuming that $J_1(r)$ is nonincreasing; the required modification for the alternate hypothesis will be self-evident. Define h by

$$h(y)^{pp'}=\int_y^\infty v(s)^{-p'}ds=J_1(y)^{p'}\left(\int_y^\infty K(s,y)^q u(s)^q ds\right)^{-p'/q}.$$

Then by Hölder's inequality

$$(Kf)(x)\ge\left(\int_{-\infty}^{x}\left[v(y)f(y)h(y)\right]^p K(x,y)^p dy\right)^{1/p}\left(\int_{-\infty}^{x}\left[v(y)h(y)\right]^{-p'}dy\right)^{1/p'},$$

and on integrating it follows that the second integral is

$$\int_{-\infty}^{x} v(y)^{-p'}\left(\int_{y}^{\infty} v(s)^{-p'}ds\right)^{-1/p}dy \le (-p')\left(\int_{x}^{\infty} v(s)^{-p'}ds\right)^{1/p'} = (-p')h(x)^{p}.$$

Denote the right side of (3.1) by I. Then using the previous estimate we obtain

$$I^{p} \ge (-p')^{p/p'}\left(\int_{-\infty}^{\infty} u(x)^{q}h(x)^{pq/p'}\left(\int_{-\infty}^{x}[v(y)f(y)h(y)]^{p}K(x,y)^{p}dy\right)^{q/p}dx\right)^{p/q}.$$

But since $q/p \le 1$, Minkowski's inequality (2.5) shows that

$$I^{p} \ge (-p')^{p/p'}\left(\int_{-\infty}^{\infty}[v(y)f(y)h(y)]^{p}\left(\int_{y}^{\infty}K(x,y)^{q}u(x)^{q}h(x)^{pq/p'}dx\right)^{p/q}dy\right)$$

$$= (-p')^{p/p'}\left(\int_{-\infty}^{\infty}[v(y)f(y)h(y)]^{p} \cdot \left(\int_{y}^{\infty}K(x,y)^{q}u(x)^{q}\left(\int_{x}^{\infty}v(s)^{-p'}ds\right)^{q/(p'p')}dx\right)^{p/q}dy\right)$$

$$= (-p')^{p/p'}\left(\int_{-\infty}^{\infty}[v(y)f(y)h(y)]^{p} \cdot \left(\int_{y}^{\infty}K(x,y)^{q}u(x)^{q}J_{1}(x)^{q/p'}\left(\int_{x}^{\infty}K(s,x)^{q}u(s)^{q}ds\right)^{-1/p'}dx\right)^{p/q}dy\right).$$

But since J_1 is nonincreasing and $K(s,x) \ge K(s,y)$ for $y<x$, we obtain on integrating

$$I^{p} \ge (-p')^{p/p'}p^{p/q}\int_{-\infty}^{\infty}[v(y)f(y)h(y)]^{p}J_{1}(y)^{p/p'}\left(\int_{y}^{\infty}K(s,y)^{q}u(s)^{q}ds\right)^{1/q}dy$$

$$= (-p')^{p/p'}p^{p/q}\int_{-\infty}^{\infty}[f(y)v(y)h(y)]^{p}J_{1}(y)^{1+p/p'}h(y)^{-p}dy$$

$$\ge (-p')^{p/p'}p^{p/q}B^{p}\int_{-\infty}^{\infty}[f(y)v(y)]^{p}dy,$$

which yields the result of part (a).

The proof of part (b) is similar to that of (a) except that now h is defined by

$$h(y)^{pp'} = \int_{-\infty}^{y} v(s)^{-p'}ds = K_{1}(y)^{p'}\left(\int_{-\infty}^{y}[K(y,s)u(s)]^{q}ds\right)^{-p'/q}.$$

The details are omitted.

Now we consider the case for negative indices.

THEOREM 3.2. *Let $K(x,y) \ge 0$ be defined in Δ and suppose $q \le p < 0$.*

(a) *If $K(x,y)$ is nonincreasing in x and (u,v) satisfy $\inf_r K_0(r) \equiv B > 0$, with $K_0(r)$ either bounded above or nondecreasing, then* (3.1) *holds for all K-admissible f.*

(b) *If $K(x,y)$ is nondecreasing in y and (u,v) satisfy $\inf_r J_0(r) \equiv B > 0$, with $J_0(r)$ either bounded above or nonincreasing, then* (3.2) *holds for all K*-admissible f.*

Proof. Part (a) is the dual of Theorem 3.1(b); part (b) is the dual of Theorem 3.1 (a). Since the proof is analogous to that of Theorem 2.2, the details are omitted.

The change of variable $x=\log t$, $y=\log s$ in Theorems 3.1 and 3.2 leads immediately to the analogous results for the half line $(0,\infty)$ rather that $(-\infty,\infty)$. If this is done and $K(x,y)$ is set equal to one, we obtain some of the results in [2].

4. The discrete case. In this section we state the discrete analogues for some of the integral inequalities proved in the preceding sections. We shall state only the discrete versions of Theorems 2.1, 2.2 and 3.1 valid for sequences on Z; the corresponding versions valid for sequences on Z^+ generalize certain inequalities of Leindler, for if we take $K(m,n)$ equal to one and $p=q$ we obtain [5, (1′), (2′), (3′) and (4′)].

THEOREM 4.1. *Let $\{K(m,n)\}$ be a nonnegative double sequence defined in $D=\{(m,n)\in Z\times Z: n\le m\}$, such that $K(m,n)$ is nonincreasing in m and nondecreasing in n. If $1\le p\le q\le\infty$ and $\{u_n\}$, $\{v_n\}$ are nonnegative sequences such that for some β, $0\le\beta\le 1$, and all integers r*

$$\left(\sum_{n=r}^{\infty} K(n,r)^{\beta q}u_n^q\right)^{1/q}\left(\sum_{n=-\infty}^{r} K(r,n)^{(1-\beta)p'}v_n^{-p'}\right)^{1/p'}\le C<\infty, \tag{4.1}$$

then for all sequences $\{a_n\}$

$$\left(\sum_{n=-\infty}^{\beta}\left|u_n\sum_{k=-\infty}^{n}K(n,k)a_k\right|^q\right)^{1/q}\le AC\left(\sum_{n=-\infty}^{\infty}|v_na_n|^p\right)^{1/p}. \tag{4.2}$$

In case $p=1$, the second sum in (4.1) *is replaced in the usual way by the supremum for $n\le r$.*

The dual of Theorem 4.1 is the following:

THEOREM 4.2. *Let $\{K(m,n)\}$ be a nonnegative double sequence defined in $D=\{(m,n)\in Z\times Z: n\le m\}$ such that $K(m,n)$ is nonincreasing in m and nondecreasing in n. If $1\le p\le q\le\infty$ and $\{u_n\}$, $\{v_n\}$ are nonnegative sequences such that for some β, $0\le\beta\le 1$, and all integers r*

$$\left(\sum_{n=-\infty}^{r} K(r,n)^{\beta q}u_n^q\right)^{1/q}\left(\sum_{n=r}^{\infty} K(n,r)^{(1-\beta)p'}v_n^{-p'}\right)^{1/p'}\le C^*<\infty, \tag{4.3}$$

then for all sequences $\{a_n\}$

$$\left(\sum_{n=-\infty}^{\infty}\left|u_n\sum_{k=n}^{\infty}K(k,n)a_k\right|^q\right)^{1/q}\le AC^*\left(\sum_{n=-\infty}^{\infty}|v_na_n|^p\right)^{1/p}. \tag{4.4}$$

The analogue of Theorem 3.1 is as follows.

THEOREM 4.3. *Let $K(m,n)$ be a nonnegative sequence defined on $D=\{(m,n)\in Z\times Z: n\le m\}$ and suppose $0<q\le p<1$.*

(a) *If $K(m,n)$ is nondecreasing in n and $\{u_k\}\ge 0$, $\{v_k\}\ge 0$ are such that*

$$J_1(r)\equiv\left(\sum_{m=r}^{\infty}\left[K(m,r)u_m\right]^q\right)^{1/q}\left(\sum_{m=r}^{\infty}v_m^{-p'}\right)^{1/p'},\qquad r\in Z,$$

satisfies $\inf_r J_1(r)\equiv B>0$, and $J_1(r)$ is either bounded above or nonincreasing, then

$$\left(\sum_{n=-\infty}^{\infty}\left[v_na_n\right]^p\right)^{1/p}\le C\left(\sum_{n=-\infty}^{\infty}\left[u_n\sum_{k=-\infty}^{n}K(n,k)a_k\right]^q\right)^{1/q} \tag{4.5}$$

holds for some constant $C>0$ and all nonnegative $\{a_k\}$ for which $\sum_{k=-\infty}^{n}K(n,k)a_k$ is finite for all n.

(b) *If* $K(m,n)$ *is nonincreasing in* m *and* $\{u_k\}\geq 0, \{v_k\}\geq 0$ *are such that*

$$K_1(r)\equiv\left(\sum_{n=-\infty}^{r}[K(r,n)u_n]^q\right)^{1/q}\left(\sum_{n=-\infty}^{r}v_n^{-p'}\right)^{1/p'}, \qquad r\in Z,$$

satisfies $\inf_r K_1(r)\equiv B>0$, *and* $K_1(r)$ *is either bounded above or nondecreasing, then*

$$\left(\sum_{n=-\infty}^{\infty}[v_n a_n]^p\right)^{1/p}\leq C\left(\sum_{n=-\infty}^{\infty}\left[u_n\sum_{k=n}^{\infty}K(k,n)a_k\right]^q\right)^{1/q} \tag{4.6}$$

holds for some constant $C>0$ *and all* $\{a_k\}\geq 0$ *for which* $\sum_{k=n}^{\infty}K(k,n)a_k$ *is finite for all* n.

The proofs of these theorems follow closely those of their integral analogues; thus the proof of Theorem 4.1 is parallel to that of Theorem 2.1, except that now, for example, the exact integration that led to (2.7) is replaced by an appeal to the easily derived inequality

$$\sum_{-\infty}^{N}c_n\left(\sum_{-\infty}^{n}c_k\right)^{-1/r}\leq r'\left(\sum_{-\infty}^{N}c_k\right)^{1/r'},$$

valid for nonnegative sequences $\{c_k\}$, $1<r<\infty$. The inequality

$$\sum_{N}^{\infty}c_n\left(\sum_{k=n}^{\infty}c_k\right)^{-1/r'}\leq r\left(\sum_{N}^{\infty}c_k\right)^{1/r}$$

is used at a later stage to complete the proof. Similarly, the elementary inequalities

$$\sum_{-\infty}^{N}c_n\left(\sum_{n}^{\infty}c_k\right)^{-1/r}\leq(1-r')\left(\sum_{N}^{\infty}c_k\right)^{1/r'},$$

$$\sum_{N}^{\infty}c_n\left(\sum_{-\infty}^{n}c_k\right)^{-1/r}\leq(1-r')\left(\sum_{-\infty}^{N}c_k\right)^{1/r'},$$

$$\sum_{N}^{\infty}c_n\left(\sum_{n}^{\infty}c_k\right)^{-1/r'}\geq r\left(\sum_{N}^{\infty}c_k\right)^{1/r},$$

$$\sum_{-\infty}^{N}c_n\left(\sum_{-\infty}^{n}c_k\right)^{-1/r'}\geq r\left(\sum_{-\infty}^{N}c_k\right)^{1/r},$$

valid for $0<r<1$ and nonnegative sequences $\{c_k\}$ are used in the course of proving the remaining theorems. We omit the details.

REFERENCES

[1] K. F. ANDERSEN AND B. MUCKENHOUPT, *Weighted weak type Hardy inequalities with applications to Hilbert transforms and maximal functions*, Studia Math., 72 (1981), pp. 9–26.

[2] P. R. BEESACK AND H. P. HEINIG, *Hardy's inequalities with indices less than* 1, Proc. Amer. Math. Soc., 83 (1981), pp. 532–536.

[3] J. S. BRADLEY, *Hardy inequalities with mixed norms*, Canad. Math. Bull., 21 (1978), pp. 405–408.

[4] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge Univ. Press, Cambridge, 1959.

[5] L. LEINDLER, *Generalization of inequalities of Hardy and Littlewood*, Acta Sci. Math., 31 (1970), pp. 279–285.

[6] B. MUCKENHOUPT, *Hardy's inequality with weights*, Studia Math., 44 (1972), pp. 31–38.

SIAM J. MATH. ANAL.
Vol. 14, No. 4, July 1983

ERRATUM: HILBERT TRANSFORM OF A FUNCTION HAVING A BOUNDED INTEGRAL AND A BOUNDED DERIVATIVE*

B. F. LOGAN†

The condition

$$f \text{ is differentiable almost everywhere with } |f'(x)| \leq m$$

should be replaced by

$$f \text{ is the integral of a bounded function } f', \text{ satisfying } |f'(x)| \leq m \ (-\infty < x < \infty).$$

* This Journal, 14 (1983), pp. 247–248.

† Bell Laboratories, Murray Hill, New Jersey 07974.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0001 $01.25/0

SUSTAINED RESONANCE FOR A NONLINEAR SYSTEM WITH SLOWLY VARYING COEFFICIENTS*

CLARK ROBINSON†

Abstract. J. Kevorkian [SIAM J. Appl. Math., 20 (1971), pp. 364–373; 26 (1974), pp. 638–669] studied resonance for a spinning reentry vehicle using a model system of ordinary differential equations with slowly varying coefficients. He and L. Lewin [SIAM J. Appl. Math., 35 (1978), pp. 738–754] gave formal multiple-time-scale expansions and numerical results to give a description of a mechanism for capture in sustained resonance. J. Sanders [SIAM J. Math. Anal., 10 (1979), pp. 1220–1243] studied these equations more rigorously using the method of averaging, but still could not prove the existence of sustained resonance. This paper continues the study of these equations using higher order averaging and the Melnikov method and shows rigorously that capture in sustained resonance does take place for some initial conditions. The Melnikov method measures the opening of a saddle connection for a small perturbation in terms of an integral. Since it has usually been applied to perturbations which depend periodically on time, the derivation of the integral for systems with slowly varying coefficients is included.

AMS-MOS subject classification (1980). Primary 34C35, 70K30

Key words. resonance, capture in resonance, reentry vehicle

1. Introduction. The system of ordinary differential equations with slowly varying coefficients studied in this paper were introduced by J. Kevorkian to model a spinning reentry vehicle [7]. (See §2 for the equations.) For a large set of initial conditions the pitch frequency becomes equal to the roll frequency, i.e. the system becomes in resonance. Most of these initial conditions lead to motion that passes through resonance, but a small set of initial conditions lead to capture in sustained resonance (on at least a time scale of $1/\varepsilon$).

J. Kevorkian studied these equations formally using multiple-time-scale expansions [6], [7], [9]. The paper with L. Lewin gives numerical results as well as a description of the capture in sustained resonance, although "no attempt is made to provide rigorous proofs" [9].

J. Sanders used the method of averaging to make a more rigorous study [14]. He studied both passage through resonance and capture in sustained resonance. He stated explicitly that he could not prove resonance is sustained on a time scale of $1/\varepsilon$ when the natural time scale is $\varepsilon^{-1/2}$. (We learned of these equations through his work.)

In this paper, we prove rigorously that capture in sustained resonance does occur. Both J. Kevorkian and J. Sanders give solutions as functions of time; we do not do this but do show the mechanism for capture in sustained resonance, using the method of higher order averaging and the Melnikov method. By using systematic higher order averaging, we include all the terms of second order in ε. The paper by L. Lewin and J. Kevorkian included one term in their description but not all the terms. J. Sanders used the method of averaging but dropped all but the lowest order terms when he formed his inner expansion. We use his averaging results but include the next order terms. See the Addendum for a comparison with the recent papers of W. Kath [16] and R. Haberman [17].

*Received by the editors September 23, 1981, and in revised form August 30, 1982. This research was partially supported by the National Science Foundation under grant MCS 81-02177.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

The Melnikov method is usually applied to periodic time perturbations of a system with a saddle connection ([10], [5a], [5b] or [1]). The Melnikov integral measures the infinitesimal separation of the stable and unstable manifolds of the saddle point as a function of the small parameter ε. We show that this method is also applicable to systems where the coefficients vary slowly with time. We show that the integral is positive for the model equations for the roll/pitch resonance problem, and so there is an opening between the stable and unstable manifolds through which orbits can be captured in sustained resonance.

I initially became involved in studying resonance problems through collaboration with J. Murdock on a related equation [13], [11]. In the equations of the earlier paper the coefficients were constant but there was a damping term. This made it possible to use a Lyapunov function argument rather than the Melnikov integral used here.

In conclusion, we mention the early paper of W. Kyner on capture in resonance that uses the separation of the separatrix solutions of a saddle point to measure the probability of the capture set [8].

2. Equations and statement of results. After describing the equations which apply to the roll, pitch, and yaw of a reentry vehicle, J. Kevorkian introduced the simpler model equations which he studied [7]:

$$\ddot{q}=-(p^2+u^2)q, \quad \dot{p}=\varepsilon u^2 q\sin\psi, \quad \dot{\psi}=2^{1/2}p, \quad \dot{u}=\tfrac{1}{2}\varepsilon u, \tag{2.1}$$

where q is the pitch angle, ψ is the roll angle, p is the roll rate, u is the natural pitch frequency when roll is not present, and the small parameter ε is a dimensionless quantity related to the change of atmospheric density at high altitude and also the distance of the center of mass from the long axis of the vehicle.

Following J. Sanders [14] we put these equations in a standard form by letting w,ξ,x and θ satisfy

$$q=w\sin\xi, \quad \dot{q}=w(p^2+u^2)^{1/2}\cos\xi, \quad x=p/u, \quad \theta=\psi-\xi.$$

(These varibles differ slightly from [14]: ξ is his $\xi+\phi$, and we take the negative of his θ.) Expanding the resulting differential equations in a finite Fourier series yields the equations

$$\dot{\theta}=2^{1/2}ux-u(1+x^2)^{1/2}$$
$$-\varepsilon(1+x^2)^{-1}\{\tfrac{1}{4}\sin 2\xi+\tfrac{1}{8}uwx[\sin(\xi+\psi)-\sin\theta+\sin(3\xi-\psi)-\sin(3\xi+\psi)]\},$$
$$\dot{x}=\tfrac{1}{2}\varepsilon\{uw\cos\theta-uw\cos(\xi+\psi)-x\},$$
$$\dot{w}=-\varepsilon w(1+x^2)^{-1}\{\tfrac{1}{4}+\tfrac{1}{4}\cos 2\xi+\tfrac{1}{8}uwx[\cos(3\xi-\psi)-\cos(3\xi+\psi)+\cos\theta-\cos(\xi+\psi)]\},$$
$$\dot{\psi}=2^{1/2}xu,$$
$$\dot{u}=\tfrac{1}{2}\varepsilon u.$$

Among the five angles, 2ξ, $\psi+\xi$, $\psi-\xi$, $3\xi+\psi$, $3\xi-\psi$, only the angle $\theta=\psi-\xi$ varies slowly and this occurs for $x=1$. All the Fourier terms with fast varying angles average to zero. By the method of higher order averaging [12], there exists a change of variables near $x=1$,

$$(\phi,z,y,\eta,u)=(\theta,x,w,\psi,u)+O(\varepsilon),$$

so the equations eliminate the terms which average to zero and they become

$$\dot{\phi}=2^{1/2}uz-u(1+z^2)^{1/2}+\tfrac{1}{8}\varepsilon(1+z^2)^{-1}uyz\sin\phi+O(\varepsilon^2),$$
$$\dot{z}=\tfrac{1}{2}\varepsilon uy\cos\phi-\tfrac{1}{2}\varepsilon z+O(\varepsilon^2),$$
$$\dot{y}=-\varepsilon y(1+z^2)^{-1}\{\tfrac{1}{4}+\tfrac{1}{8}uyz\cos\phi\}+O(\varepsilon^2),$$
$$\dot{\eta}=2^{1/2}zu+O(\varepsilon^2),$$
$$\dot{u}=\tfrac{1}{2}\varepsilon u.$$

Since η does not appear in the equations, except for the $O(\varepsilon^2)$ terms, and since we show below that the terms of order ε determine the behavior, we drop this variable from further consideration.

In the resonant band about $z=1$ (or $x=1$, or $p=u$), letting $z=1+\mu\zeta$, $\tau=\mu t$ and $(')$ be $d/d\tau$, the equations become

$$\phi'=(2^{1/2}/2)u\zeta-\mu(2^{1/2}/8)u\zeta^2+\mu(\tfrac{1}{16})uy\sin\phi+O(\mu^2),$$
$$\zeta'=\tfrac{1}{2}uy\cos\phi-\tfrac{1}{2}-u\zeta/2+O(\mu^2),$$

or as a second order equation in ϕ letting $v=\phi'$ and $h=uy\cos\phi-1$,

$$\begin{aligned}\phi'&=v,\\ v'&=(2^{1/2}/4)uh-\mu(\tfrac{3}{16})hv+\mu(\tfrac{1}{16})v+O(\mu^2),\\ y'&=-\mu(\tfrac{1}{8})y-\mu(\tfrac{1}{16})uy^2\cos\phi+O(\mu^2),\\ u'&=\mu(\tfrac{1}{2})u.\end{aligned}\tag{2.2}$$

The inner equations of [14, 8.19] are the same as these but do not include the terms of order μ, except for the u' equation.

Equations (2.2) are the ones we study. The averaged equations are obtained by dropping the terms of order μ^2. Numerical integration of these averaged equations leads to capture in sustained resonance for the initial conditions

$$\phi_0=-2.5,\quad y_0=1.5,\quad u_0=1,\quad \varepsilon=0.0001,\quad 1.16\le v_0\le 1.24.$$

See Fig. 1. Compare with [9, Fig. 7].

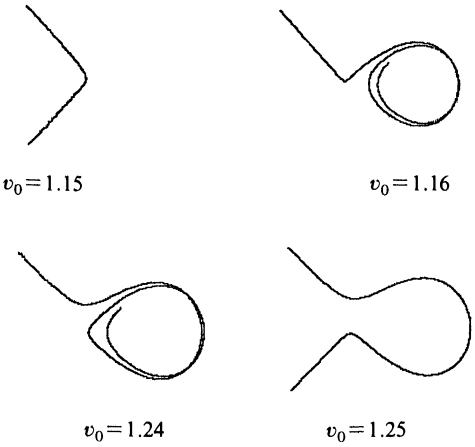


FIG. 1.

To understand Fig. 1, notice that for $\mu=0$, y and u are constants, and there are fixed points at $v=0$ and $\cos\phi=1/uy$. One of the fixed points is a hyperbolic saddle with a saddle connection and the other is an elliptic fixed point. See Fig. 2.

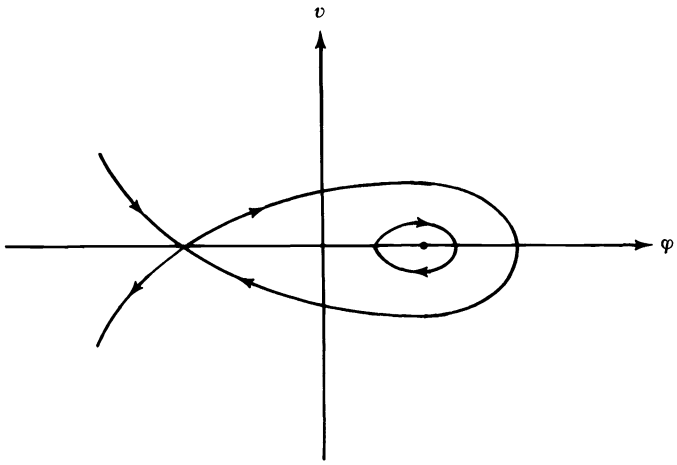


FIG. 2. *Flow in* (ϕ,v)*-space for* $\mu=0$.

We prove that for small μ and $1<uy$, the unstable manifold comes inside the stable manifold, leaving an opening in between to capture an orbit, so v is bounded by a constant C_1 for $0\le t\le C_2/\varepsilon$. The reason the time interval is only $O(1/\varepsilon)$ is that the bound $\varepsilon_0(u,y)$ is not proved to be uniform in u and y, and it takes time on the order of ε^{-1} to leave a compact region in the set $\{1\le uy\}$.

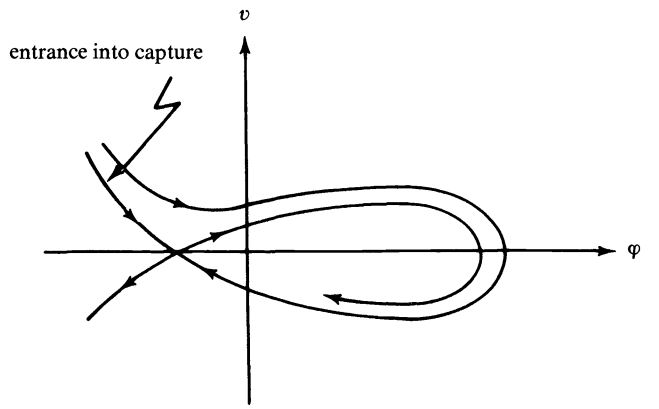


FIG. 3. *Flow in* (ϕ,v)*-space for* $\mu>0$.

If an orbit has $|v(t)|\le C_1$ for $0\le t\le C_2/\varepsilon$, then $v\approx u\zeta$, so

$$|u(t)-p(t)|=|u(t)(1-x(t))|=|u(t)\varepsilon^{1/2}\zeta(t)|\le\varepsilon^{1/2}C_1$$

for $0 \le t \le C_2/\varepsilon$. Thus for such an orbit the roll rate $p(t)$ grows at an exponential rate like $u(t)$ on a time interval $0 \le t \le C_2/\varepsilon$. See [9, Fig. 5]. This motivates the definition of sustained resonance given in [9, p. 748]: an orbit is in *sustained resonance* if $|u(t)-p(t)| = O(\varepsilon^{1/2})$ for a time interval $0 \le t \le C_1/\varepsilon$. We can now state our main theorem.

THEOREM 2.3. *For $1<u_0y_0$ there exists $\varepsilon_0(u_0,y_0)>0$ such that for $0<\varepsilon<\varepsilon_0(u_0,y_0)$, there are some orbits for* (2.2) (*or* (2.1)) *which start outside the resonant band and are captured in sustained resonance for a time interval of at least $O(1/\varepsilon)$. For these capture orbits, $|u(t)-p(t)| \le C_1\varepsilon^{1/2}$ for $0 \le t \le C_2/\varepsilon$, so the roll rate builds up exponentially on this time interval.* (*Many other orbits enter the resonant band and leave after a short time.*)

As mentioned above, for $\mu=0$ (or $\varepsilon=0$) there is a hyperbolic fixed point for every fixed set of parameters u and y with $1<uy$. Because for $\mu>0$ the parameters u and y vary slowly, there is not an actual fixed point, but there is an invariant set in (ϕ,v,u,y)-space with codimension one stable and unstable manifolds. (See §3 for details.)

The distance between the stable and unstable manifolds for $\mu>0$ is measured by the Melnikov integral, $\Delta_1(0,u,y)$, defined in §3. This integral measures the effect of the next order terms in μ on a saddle connection. The fact that $\Delta_1(0,u,y)>0$ for $1<uy$, implies that for small enough μ the unstable manifold comes inside the stable manifold, leaving an opening between to capture an orbit in resonance. (See Fig. 3.) The Melnikov method is usually applied to time periodic perturbations, so we show in §3 that it is applicable to systems with slowly varying coefficients. At the end of this section we apply it to a simple but general forced pendulum problem with slowly varying coefficients. In §4, the Melnikov integral for (2.2) is derived and proved to be positive for $1<uy$.

Finally in §5, the separation of the stable and unstable manifolds predicted by the Melnikov integral is compared with results from numerical integration of the differential equations (2.2) themselves.

3. Melnikov method for slowly varying coefficients. The usual Melnikov method applies to ordinary differential equations with periodic dependence on time ([10], [5a], [15], or [1]). In this section we show how the method applies when the coefficients vary slowly with time. More specifically, consider the equations

$$
\begin{aligned}
\theta' &= p_0(\theta,v,\mathbf{u}) + \varepsilon p_1(\theta,v,\mathbf{u}) + O(\varepsilon^2),\\
v' &= q_0(\theta,v,\mathbf{u}) + \varepsilon q_1(\theta,v,\mathbf{u}) + O(\varepsilon^2),\\
\mathbf{u}' &= \varepsilon r(\theta,v,\mathbf{u}) + O(\varepsilon^2),
\end{aligned} \tag{3.1}
$$

where θ and v are scalars and $\mathbf{u}$ is allowed to be a vector quantity. Letting $\mathbf{x}=(\theta,v)$, these equations can be written as

$$
\begin{aligned}
\mathbf{x}' &= f_0(\mathbf{x},\mathbf{u}) + \varepsilon f_1(\mathbf{x},\mathbf{u}) + O(\varepsilon^2),\\
\mathbf{u}' &= \varepsilon r(\mathbf{x},\mathbf{u}) + O(\varepsilon^2).
\end{aligned} \tag{3.1$'$}
$$

Saddle connection assumption 3.2. The equations for $\varepsilon=0$ and for $\mathbf{u}_0$ in a bounded set U are assumed to have a hyperbolic saddle point $\mathbf{z}(\mathbf{u}_0,0)$ with a saddle connection with solution $\mathbf{x}_0(t,\mathbf{u}_0)$.

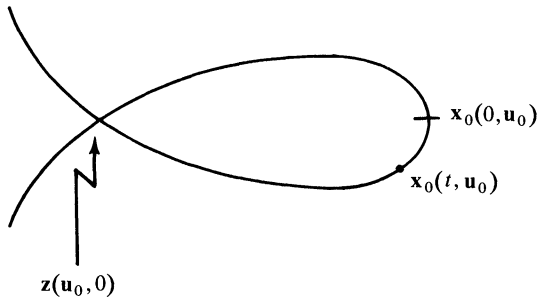


FIG. 4.

PROPOSITION 3.3. *Assume equations* (3.1) *satisfy the saddle connection assumption. Then for* $0 \leq \varepsilon \ll 1$ *there is a normally hyperbolic invariant set,*

$$M_\varepsilon = \{(\mathbf{x},\mathbf{u})\colon \mathbf{x}=\mathbf{z}(\mathbf{u},\varepsilon) \text{ with } \mathbf{u}\in U\},$$

where $\mathbf{z}$ *is a* C^1 *function of* $\mathbf{u}$, *and* ε *and* $\mathbf{z}(\mathbf{u},0)$ *are the saddle fixed points. (Note the set* U *is not invariant under the flow of* (3.1), *so* M_ε *is an invariant set in the weaker sense that a solution can leave only by its* $\mathbf{u}$ *variable crossing the boundary of* U.) *Motion on* M_ε *is slow, i.e., on the order of* ε. *Moreover, the invariant set* M_ε *possesses stable and unstable manifolds,* $W^s(\varepsilon)$ *and* $W^u(\varepsilon)$ *of points which approach* M_ε *exponentially fast as* t *goes to* ∞ *or* $-\infty$, *respectively. These manifolds are* C^1 *functions of* $\mathbf{u}$ *and* ε, *and so they are* C^1 *close to the saddle connection on compact subsets.*

Proof. The existence of M_ε, $W^s(\varepsilon)$ and $W^u(\varepsilon)$ follows from the persistence of normally hyperbolic invariant sets and their stable and unstable manifolds ([4, Thm. 4.1] or [2, Thm. 3]). The usual theorem requires a compact set (or uniform estimates on derivatives on all of a Euclidean space), so it is necessary to patch $U \subset U^*$ where U^* is a compact manifold (e.g. a sphere) and extend (3.1) to a neighborhood of a hyperbolic set

$$\{(\mathbf{x},\mathbf{u})\colon \mathbf{u}\in U^*,\ \mathbf{x}=\mathbf{z}(\mathbf{u},0)\}.$$

The fact that motion is slow on M_ε follows because it is a graph over the $\mathbf{u}$ variables.

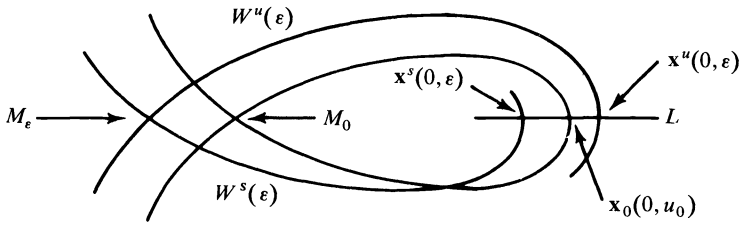


FIG. 5. $\mathbf{x}$-*space with* $\mathbf{u}_0$ *fixed.*

The Melnikov method measures the separation of $W^s(\varepsilon)$ and $W^u(\varepsilon)$ for $0<\varepsilon\ll 1$. Fix $\mathbf{u}_0$. Let L be the line through $\mathbf{x}_0(0,u_0)=\mathbf{x}_0(0)$, in $\mathbf{x}$-space, which is perpendicular to $f_0(\mathbf{x}_0(0),\mathbf{u}_0)$. Let $\mathbf{x}^s(t,\varepsilon)$ (resp. $\mathbf{x}^u(t,\varepsilon)$) be the solution in $W^s(\varepsilon)$ (resp. $W^u(\varepsilon)$) crossing L at $t=0$ and also with $\mathbf{u}=\mathbf{u}_0$ at $t=0$. Let $\mathbf{x}_0(t)=\mathbf{x}_0(t,\mathbf{u}_0)$, and $f_0(\mathbf{x}_0(t))=f_0(\mathbf{x}_0(t),\mathbf{u}_0)$.

Letting the wedge product represent the scalar cross product in the plane, define

$$\Delta(t,\mathbf{u}_0,\varepsilon)=[\mathbf{x}^u(t,\varepsilon)-\mathbf{x}^s(t,\varepsilon)]\wedge f_0(\mathbf{x}_0(t))$$

and

$$\Delta_1(t,\mathbf{u}_0)=\frac{\partial}{\partial\varepsilon}\Delta(t,\mathbf{u}_0,\varepsilon)\Big|_{\varepsilon=0}=\left[\frac{\partial}{\partial\varepsilon}\mathbf{x}^u(t,\varepsilon)-\frac{\partial}{\partial\varepsilon}\mathbf{x}^s(t,\varepsilon)\right]\wedge f_0(\mathbf{x}_0(t)).$$

The function $\Delta_1(0,\mathbf{u}_0)$ measures the infinitesimal separation of the stable and unstable manifolds as a function of ε. (J. Sanders has pointed out that Δ is not uniquely defined but depends on L. The use of Δ_1 eliminates this difficulty.) Notice here and below that ε affects the solution of $\mathbf{x}^s$ and $\mathbf{x}^u$ directly through the dependence of $f(\mathbf{x},\mathbf{u},\varepsilon)$ on ε, and indirectly through the dependence of $\mathbf{u}$ at time t on ε. The following result is the heart of the Melnikov method.

PROPOSITION 3.4. *Assume equations* (3.1) *satisfy the saddle connection assumption, and that* f_0 *is divergence free as a function of* $\mathbf{x}$, *i.e.* $\operatorname{tr} Df_0(\mathbf{x},\mathbf{u})\equiv 0$, *where* D *is the derivative with respect to* $\mathbf{x}$. *Then*

$$\Delta_1(0,\mathbf{u}_0)=\int_{-\infty}^{\infty}\left\{f_1(\mathbf{x}_0(t),\mathbf{u}_0)+\frac{\partial f_0}{\partial\mathbf{u}}(\mathbf{x}_0(t),\mathbf{u}_0)\frac{\partial\mathbf{u}}{\partial\varepsilon}\right\}\wedge f_0(\mathbf{x}_0(t),\mathbf{u}_0)\,dt$$

where $\partial\mathbf{u}/\partial\varepsilon$ *satisfies*

$$\left(\frac{\partial\mathbf{u}}{\partial\varepsilon}\right)'=r(\mathbf{x}_0(t),\mathbf{u}_0)\quad\text{and}\quad\frac{\partial\mathbf{u}}{\partial\varepsilon}=0\quad\text{at } t=0.$$

Using p *and* q, *the integral can be written*

$$\Delta_1(0,\mathbf{u}_0)=\int_{-\infty}^{\infty}q_0p_1-p_0q_1+\left(q_0\frac{\partial p_0}{\partial\mathbf{u}}-p_0\frac{\partial q_0}{\partial\mathbf{u}}\right)\cdot\frac{\partial\mathbf{u}}{\partial\varepsilon}\,dt.$$

Remark. If f_0 is not divergence free, then the integral changes as in [5].

Proof. Take all partial derivatives with respect to ε at $\varepsilon=0$. Let D be the derivative (Jacobian matrix) with respect to $\mathbf{x}$. Let

$$\Delta_1^u(t)=\frac{\partial}{\partial\varepsilon}\mathbf{x}^u(t,\varepsilon)\wedge f_0(\mathbf{x}_0(t)),$$

$$\Delta_1^s(t)=\frac{\partial}{\partial\varepsilon}\mathbf{x}^s(t,\varepsilon)\wedge f_0(\mathbf{x}_0(t)),$$

so $\Delta_1(t,\mathbf{u}_0)=\Delta_1^u(t)-\Delta_1^s(t)$. Taking the derivative of $\Delta_1^s(t)$ with respect to t gives

$$(\Delta_1^s)'=\left(\frac{\partial\mathbf{x}^s}{\partial\varepsilon}\right)'\wedge f_0+\frac{\partial\mathbf{x}^s}{\partial\varepsilon}\wedge Df_0(\mathbf{x}_0(t))f_0.$$

By theorems on dependence of solutions on parameters [3, Thm. V.3.1]

$$\left(\frac{\partial\mathbf{x}^s}{\partial\varepsilon}\right)'=Df_0\cdot\frac{\partial\mathbf{x}^s}{\partial\varepsilon}+f_1+\frac{\partial f_0}{\partial\mathbf{u}}\frac{\partial\mathbf{u}}{\partial\varepsilon},$$

all evaluated at $\mathbf{x}=\mathbf{x}_0(t)$, $\mathbf{u}=\mathbf{u}_0$. Therefore

$$(\Delta_1^s)'=\left\{f_1+\left(\frac{\partial f_0}{\partial\mathbf{u}}\right)\cdot\left(\frac{\partial\mathbf{u}}{\partial\varepsilon}\right)\right\}\wedge f_0+\left\{Df_0\cdot\left(\frac{\partial\mathbf{x}^s}{\partial\varepsilon}\right)\wedge f_0+\left(\frac{\partial\mathbf{x}^s}{\partial\varepsilon}\right)\wedge Df_0f_0\right\}.$$

The second term equals

$$(\operatorname{tr} Df_0)\left\{\frac{\partial \mathbf{x}^s}{\partial \varepsilon}\wedge f_0\right\}=0,$$

because f_0 is assumed divergence free. The first term is like the integrand of the statement of the result. Letting $F=f_1+(\partial f_0/\partial \mathbf{u})(\partial \mathbf{u}/\partial \varepsilon)$ gives

$$(\Delta_1^s)'=F\wedge f_0,$$

so

$$\Delta_1^s(T)-\Delta_1^s(0)=\int_0^T F\wedge f_0\,dt,$$
$$-\Delta_1^s(0)=-\Delta_1^s(T)+\int_0^T F\wedge f_0\,dt.$$

Below it is shown that

$$\lim_{T\to\infty}\Delta_1^s(T)=0,$$

so

$$-\Delta_1^s(0)=\int_0^\infty F\wedge f_0\,dt.$$

Similarly for Δ_1^u, letting $T\to-\infty$

$$\Delta_1^u(0)=\Delta_1^u(T)+\int_T^0 F\wedge f_0\,dt=\int_{-\infty}^0 F\wedge f_0\,dt,$$

and

$$\Delta_1(0,\mathbf{u}_0)=\int_{-\infty}^\infty F\wedge f_0\,dt.$$

Finally before checking the limits, by the theorem on dependence of solutions on parameters [3, Thm. V.3.1], $\partial \mathbf{u}/\partial \varepsilon$ satisfies

$$\left(\frac{\partial \mathbf{u}}{\partial \varepsilon}\right)'=\left(\frac{\partial}{\partial \mathbf{u}}\varepsilon r(\mathbf{x}(t),\mathbf{u}(t))\right)\frac{\partial \mathbf{u}}{\partial \varepsilon}+\frac{\partial}{\partial \varepsilon}(\varepsilon r(\mathbf{x}(t),\mathbf{u}(t))).$$

Because we are interested in this for $\varepsilon=0$, the first term vanishes, $\mathbf{u}(t)=\mathbf{u}_0$ is a constant, and $\mathbf{x}(t)$ is along the homoclinic orbit.

LEMMA 3.5.

$$\lim_{T\to\infty}\Delta_1^s(T)=0=\lim_{T\to-\infty}\Delta_1^u(T).$$

Proof. In the usual time-dependence case, this result is easy because

$$\Delta_1^s(T)=\frac{\partial \mathbf{x}^s}{\partial \varepsilon}\wedge f_0(\mathbf{x}_0(T)),$$

$f_0(\mathbf{x}_0(T))$ goes to zero exponentially fast, and $\partial \mathbf{x}^s/\partial \varepsilon$, which measures the dependence of the fixed point on ε, is bounded as T goes to infinity. In our case $\partial \mathbf{x}^s/\partial \varepsilon$ is not bounded but can grow like T. The reason it can grow is because of the $\mathbf{u}$ dependency. Remember that

$$M_\varepsilon=\{(\mathbf{x},\mathbf{u})\colon \mathbf{x}=\mathbf{z}(\mathbf{u},\varepsilon)\}.$$

Let $\mathbf{u}(T,\varepsilon)$ be the value of $\mathbf{u}$ at time T on the solution corresponding to $\mathbf{x}^s(T,\varepsilon)$. We show that $\partial\mathbf{x}^s/\partial\varepsilon$ grows like $(\partial/\partial\varepsilon)\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)$. Since $\mathbf{u}(T,\varepsilon)\sim\varepsilon T$, i.e. the solution moves a distance of εT along M_ε, $(\partial/\partial\varepsilon)\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)\sim T$ and is not bounded even though M_ε is uniformly C^1.

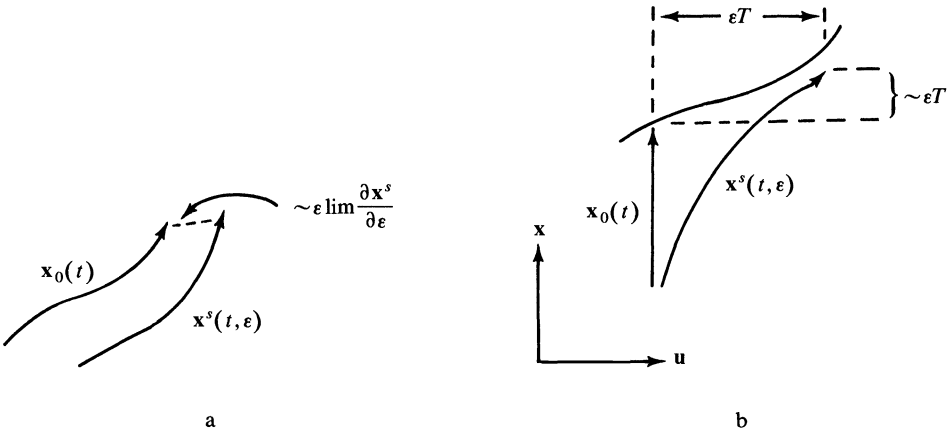


FIG. 6. a) *Usual case*, x-*space*. b) *Slowly varying coefficients*, $(\mathbf{x},\mathbf{u})$-*space*.

More precisely, as $T\to\infty$, $\mathbf{x}^s(T,\varepsilon)$ approaches $\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)$ exponentially fast. Therefore

$$\left(\frac{\partial}{\partial\varepsilon}\mathbf{x}^s\right)'-\left(\frac{\partial}{\partial\varepsilon}\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)\right)'$$

becomes small as $T\to\infty$. Then

$$\left(\frac{\partial}{\partial\varepsilon}\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)\right)'=\left(\frac{\partial\mathbf{z}}{\partial\varepsilon}\right)'+\left[\left(\frac{\partial\mathbf{z}}{\partial\mathbf{u}}\right)\left(\frac{\partial\mathbf{u}}{\partial\varepsilon}\right)\right]'.$$

Since the partials are evaluated at $\varepsilon=0$, $(\partial\mathbf{z}/\partial\varepsilon)(\mathbf{u}(T,0),0)=\partial\mathbf{z}/\partial\varepsilon(\mathbf{u}_0,0)$ has $(\partial\mathbf{z}/\partial\varepsilon)'=0$. Similarly $\frac{d}{dt}(\partial\mathbf{z}/\partial\mathbf{u})(\mathbf{u}_0,0)=0$. Therefore

$$\left(\frac{\partial}{\partial\varepsilon}\mathbf{z}(\mathbf{u}(T,\varepsilon),\varepsilon)\right)'=\left(\frac{\partial\mathbf{z}}{\partial\mathbf{u}}\right)\left(\frac{\partial\mathbf{u}}{\partial\varepsilon}\right)',$$

where $\partial\mathbf{z}/\partial\mathbf{u}$ and $(\partial\mathbf{u}/\partial\varepsilon)'=r(\mathbf{x}_0(t),\mathbf{u}_0)$ are bounded. Since $(\partial\mathbf{x}^s/\partial\varepsilon)'$ is exponentially close for large T, it too is bounded and $(\partial\mathbf{x}^s/\partial\varepsilon)(T,\varepsilon=0)$ is bounded by CT for some C. Since $f_0(\mathbf{x}_0(T))$ goes to zero exponentially fast, $\Delta_1^s(T)\to 0$ as $T\to\infty$. Similarly $\Delta_1^u(T)\to 0$ as $T\to-\infty$.

PROPOSITION 3.6. *If* $\Delta_1(0,\mathbf{u}_0)>0$ *for* $\mathbf{u}_0\in U$, *then there is capture in sustained resonance. In fact for those solutions captured, the only way they can escape resonance is for the solution to have its* $\mathbf{u}$ *value leave the region where* $\Delta_1>0$ (*or where the equations are valid*) *which takes time on the order of* $1/\varepsilon$. *This holds even with the terms* $O(\varepsilon^2)$.

Proof. The solution can not cross the manifold $W^s(\varepsilon)$ which is invariant and of codimension one in $(\mathbf{x},\mathbf{u})$-space. Capture is possible, because for $\Delta_1>0$ the stable manifold lies outside the unstable manifold leaving a gap in between where trajectories can enter into sustained resonance. See Fig. 3.

The following simple example illustrates the calculation of Δ_1. This example can be treated by other methods as was done in [9, 3.12] for the case when $f(w)=w^2$, $g(w)=w$ and $r(\varphi, v, w)=w/2$.

Example 3.7. Consider the system of equations

$$\dot{\varphi}=v, \quad \dot{v}=f(\mathbf{w})\cos\varphi-g(\mathbf{w}), \quad \dot{\mathbf{w}}=\varepsilon r(\varphi, v, \mathbf{w}).$$

Letting $f_w=(\partial f/\partial \mathbf{w})$ and $g_w=(\partial g/\partial \mathbf{w})$, assume there is a domain W, such that for $\mathbf{w}\in W$ the following five conditions are satisfied:

(i) $-1<g(\mathbf{w})/f(\mathbf{w})<1$.

(ii) $\varepsilon^{-1}df/dt(\mathbf{w})=f_w\cdot r\geq 0$ (i.e. f is nondecreasing along the saddle connection).

(iii) $-\varepsilon^{-1}(d/dt)g(\mathbf{w})/f(\mathbf{w})=f(\mathbf{w})^{-2}(-fg_w+gf_w)\cdot r\geq 0$ (i.e. f increases at least as fast as g).

(iv) One of the inequalities in (ii) and (iii) is strict at some point along the saddle connection.

(v) $f(\mathbf{w})>0$ (if $f(\mathbf{w})<0$ for all w in W then reverse all the inequalities in (ii) and (iii)).

Then, this system has a saddle connection, and $\Delta_1(0, \mathbf{w})>0$ for all $\mathbf{w}$ in W.

Proof. Condition (i) insures the existence of a hyperbolic saddle connection. Assume $f(\mathbf{w})>0$. Let s be a new time scale which solves $ds/dt=f(\mathbf{w})^{1/2}$. Letting $(')$ be d/ds and $V=vf(\mathbf{w})^{-1/2}$,

$$V'=vf^{-1}-v\tfrac{1}{2}f^{-1}f_w\cdot rf^{-1/2},$$

so the equations become

$$\begin{aligned}
\varphi' &= V,\\
V' &= \cos\varphi-g(\mathbf{w})/f(\mathbf{w})-\varepsilon\left(\tfrac{1}{2}\right)Vf(\mathbf{w})^{-1}f_w r,\\
\mathbf{w}' &= \varepsilon r\left(\varphi, Vf(\mathbf{w})^{1/2}, \mathbf{w}\right)f(\mathbf{w})^{-1/2}.
\end{aligned}$$

Take the parametrization of the saddle connection so that $V(0)=0$. Then $-sV(s)\geq 0$ along the saddle connection. The integral is then

$$\begin{aligned}
\Delta_1(0, \mathbf{w}) &= -\int_{-\infty}^{\infty} p_0 q_1\, ds-\int_{-\infty}^{\infty} p_0\frac{\partial q_0}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial \varepsilon}\, ds\\
&= \int_{-\infty}^{\infty}\frac{1}{2}V^2f^{-1}f_w\cdot r\, ds+\int_{-\infty}^{\infty}(-V)f^{-2}(-fg_w+gf_w)\cdot\frac{\partial \mathbf{w}}{\partial \varepsilon}\, ds.
\end{aligned}$$

The first integral is ≥ 0 by (ii). In the second integral f, f_w, g and g_w are constant along the integral, so

$$\frac{d}{ds}\left[(-fg_w+gf_w)\cdot\frac{\partial \mathbf{w}}{\partial \varepsilon}\right]=(-fg_w+gf_w)\cdot rf^{-1/2}\geq 0$$

by (iii), and so $(-fg_w+gf_w)$ $(\partial \mathbf{w}/\partial\varepsilon)$ has the same sign as s. Since $-V$ also has the same sign as s, the second integral is ≥ 0. By (iv) one of these two integrals is strictly positive.

4. Proof of Theorem 2.3. Using Propositions 3.3, 3.4, and 3.6, the only thing that is necessary to check is that $\Delta_1(0, u, y)>0$ for $uy>1$. The verification follows the outline of Example 3.7 and is only slightly more complicated. The power of the Melnikov

method is that it shows rigorously that the $\mu vv'$ term in (2.2) does not affect the separation of the stable and unstable manifolds, and that the terms $O(\mu^2)$ can be ignored.

As in Example 3.7, the requirement that $uy>1$ insures that there is a hyperbolic saddle point with a saddle connection for $\mu=0$. Introduce a new time scale s by solving $ds/d\tau=uy^{1/2}$. Letting $H=8^{-1/2}\cos\varphi-8^{-1/2}/(uy)$, (2.2) becomes

$$\begin{aligned}
\frac{d\varphi}{ds}&=V,\\
\frac{dV}{ds}&=H-\mu 8^{1/2}\left(\tfrac{5}{32}\right)y^{1/2}VH-\mu\left(\tfrac{11}{32}\right)u^{-1}y^{-1/2}V+O(\mu^2),\\
\frac{dy}{ds}&=-\tfrac{3}{16}y^{1/2}u^{-1}-(8^{1/2}/16)y^{3/2}H+O(\mu^2),\\
\frac{du}{ds}&=\mu\tfrac{1}{2}y^{-1/2}.
\end{aligned}$$

Then

$$\begin{aligned}
\Delta_1(0,u,y)=&-\int_{-\infty}^{\infty}p_0q_1\,ds-\int_{-\infty}^{\infty}p_0\frac{\partial q_0}{\partial u}\frac{\partial u}{\partial\mu}ds\\
&-\int_{-\infty}^{\infty}p_0\frac{\partial q_0}{\partial y}\frac{\partial y}{\partial\mu}ds.
\end{aligned}$$

The first integral equals

$$\begin{aligned}
&\int_{-\infty}^{\infty}8^{1/2}\frac{5}{32}y^{1/2}V^2V'\,ds+\int_{-\infty}^{\infty}\frac{11}{32}u^{-1}y^{-1/2}V^2\,ds\\
&\qquad=8^{1/2}\frac{5}{96}y^{1/2}V^3\Big|_{-\infty}^{\infty}+\int_{-\infty}^{\infty}\frac{11}{32}u^{-1}y^{-1/2}V^2\,ds\\
&\qquad=\int_{-\infty}^{\infty}\frac{11}{32}u^{-1}y^{-1/2}V^2\,ds.
\end{aligned}$$

The last two integrals equal

$$\int_{-\infty}^{\infty}(-V)\left[8^{-1/2}u^{-2}y^{-1}\frac{\partial u}{\partial\mu}+8^{-1/2}u^{-1}y^{-2}\frac{\partial y}{\partial\mu}\right]ds.$$

Since

$$\begin{aligned}
&\frac{d}{ds}\left[8^{-1/2}u^{-2}y^{-1}\frac{\partial u}{\partial\mu}+8^{-1/2}u^{-1}y^{-2}\frac{\partial y}{\partial\mu}\right]\\
&\qquad=8^{-1/2}u^{-2}y^{-3/2}\left(\frac{1}{2}-\frac{3}{16}\right)-\frac{1}{16}u^{-1}y^{-1/2}V',
\end{aligned}$$

it follows that

$$\begin{aligned}
&\left[8^{-1/2}u^{-2}y^{-1}\frac{\partial u}{\partial\mu}+8^{-1/2}u^{-1}y^{-2}\frac{\partial y}{\mu}\right]\\
&\qquad=8^{-1/2}\frac{5}{16}u^{-2}y^{-3/2}s-\frac{1}{16}u^{-1}y^{-1/2}V,
\end{aligned}$$

and

$$\Delta_1(0,u,y)=\int_{-\infty}^{\infty}\frac{13}{32}u^{-1}y^{-1/2}V^2\,ds$$
$$+\int_{-\infty}^{\infty}8^{-1/2}\frac{5}{16}u^{-2}y^{-3/2}s(-V)\,ds.$$

The first integral is clearly positive, and the second is positive because the saddle connection is parametrized so that $-sV\geq 0$.

5. Comparison of the Melnikov integral and numerical integration. The gap $\Delta\varphi$ between the stable and unstable manifolds as they cross the φ-axis ($v=0$) should approximately be equal to $\varepsilon^{1/2}\Delta_1/q_0$, where q_0 is the component of the vector field in the v direction. (It is multiplied by $\mu=\varepsilon^{1/2}$, because this is the small parameter in (2.2).) To compare the predicted gap with numerical integration, various initial conditions were tried for φ. The point where the solution changes from staying captured to escaping is an estimate for the stable manifold. A similar calculation backward in time gives an estimate for the unstable manifold.

The value of the parameters used are $\varepsilon=0.001$, $u=1.2$ and $y=1.4$ (these correspond approximately to the values used in [9]). Numerically calculating the integral gives $\Delta_1=3.7$. The saddle connection crosses $v=0$ at approximately $\phi=1.97$, where the vector field is 0.70. The Melnikov integral therefore predicts that the separation is 0.17. Using numerical integration starting at $v=0$, backward in time it switches from capture to escape between 1.87 and 1.88. Forward in time it switches between 2.05 and 2.06. Therefore the separation found is about 0.18. The predicted separation is therefore a good approximation of the empirical separation.

Addendum. After this paper was written, we became aware of the recent work of W. Kath on these same equations [16]. He allows a more general form of u (his ω) than $u^2=u_0^2e^{\varepsilon t}$ and gives a good explanation of how increasing du/dt causes release from capture in resonance. This addendum indicates what conditions on du/dt imply that Δ_1 is positive so that resonance persists.

If $\dot{u}=\varepsilon r(u,\sigma)$ with $r(u,\sigma)>0$ and $\sigma=\varepsilon t$ a slow time ($\sigma=\tilde{t}$ in [16]) is allowed, then $v'=8^{-1/2}u^2y\cos\varphi-8^{-1/2}2r+O(\mu)$. A necessary condition for an equilibrium to exist becomes $u^2y>2r$. Compare with [16, 2.15]. If $r(u,\sigma)=u/2$ as considered in this paper, then this agrees with the condition that $uy>1$.

With these changes, (4.1) becomes

$$\frac{d\varphi}{ds}=V,$$
$$\frac{dV}{ds}=H-\mu 8^{1/2}\left(\tfrac{5}{32}\right)y^{1/2}VH-\mu\left(\tfrac{11}{16}\right)y^{-1/2}u^{-2}Vr+O(\mu^2)$$
$$\frac{dy}{ds}=-\mu\left(\tfrac{3}{8}\right)y^{1/2}u^{-2}r-\mu\left(8^{1/2}/16\right)y^{3/2}H+O(\mu^2),$$
$$\frac{du}{ds}=\mu y^{-1/2}u^{-1}r,$$
$$\frac{d\sigma}{ds}=\mu y^{-1/2}u^{-1},$$

where $H=8^{-1/2}\cos\varphi-8^{-1/2}2y^{-1}u^{-2}r$. Then

$$\Delta_1(0,u,y)=-\int_{-\infty}^{\infty}p_0\left[q_1+\frac{\partial q_0}{\partial u}\frac{\partial u}{\partial \mu}+\frac{\partial q_0}{\partial y}\frac{\partial y}{\partial \mu}+\frac{\partial q_0}{\partial \sigma}\frac{\partial \sigma}{\partial \mu}\right]ds$$

$$=\left(\frac{13}{16}\right)y^{-1/2}u^{-2}r\int_{-\infty}^{\infty}V^2\,ds$$

$$+8^{-1/2}u^{-3}y^{-3/2}\left[-2r\frac{\partial r}{\partial u}-2\frac{\partial r}{\partial \sigma}+\left(\frac{13}{4}\right)u^{-1}r^2\right]\int_{-\infty}^{\infty}s(-V)\,ds.$$

Therefore Δ_1 is positive if

$$-2r\frac{\partial r}{\partial u}-2\frac{\partial r}{\partial \sigma}+\left(\frac{13}{4}\right)u^{-1}r^2>0,$$

or, since $du/d\sigma=r(u,\sigma)$,

$$0>\frac{\partial r}{\partial u}\frac{du}{d\sigma}u^{-13/8}+\frac{\partial r}{\partial \sigma}u^{-13/8}-\left(\frac{13}{8}\right)u^{-21/8}r^2,$$

$$0>\frac{d}{d\sigma}\left[r(u,\sigma)u^{-13/8}\right].$$

Thus a sufficient condition to allow capture in resonance (and sustain resonance) is that $\varepsilon^{-1}du/dt=du/d\sigma=r(u,\sigma)$ grows more slowly than $u^{13/8}$ (and $\frac{1}{2}u^2y>r(u,\sigma)$ to preserve the fixed point). This answers the question at the end of [16, §3].

Another paper we received after this paper was written is by R. Haberman [17]. It uses energy methods to show that capture in resonance takes place. These methods are another way of expressing the calculations of Proposition 3.4.

REFERENCES

[1] S. N. Chow, J. Hale and J. Mallet Paret, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351–373.

[2] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[3] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York/London/Sydney, 1964.

[4] M. Hirsch, C. Pugh and M. Shub, *Invariant Manifolds*, Lecture Notes in Mathematics, 583, Springer-Verlag, Berlin/Heidelberg/New York, 1977.

[5a] P. Holmes, *Averaging and chaotic motion in forced oscillations*, SIAM J. Appl. Math., 38 (1980), pp. 65–80.

[5b] P. Holmes and J. Marsden, *A partial differential equation with infinitely many periodic orbits*, Arch. Rational Mech. Anal., 76 (1981), pp. 131–166.

[6] J. Kevorkian, *Passage through resonance for a one-dimensional oscillator with slowly varying frequencies*, SIAM J. Appl. Math., 20 (1971), pp. 364–373.

[7] ———, *On a model for reentry roll resonance*, SIAM J. Appl. Math., 26 (1974), pp. 638–669.

[8] W. T. Kyner, *Passage through resonance*, in Periodic Orbits, Stability, and Resonances, Geo Giacaglia, ed., D. Reidel, Dordrecht, the Netherlands, 1970, pp. 501–514.

[9] L. Lewin and J. Kevorkian, *On the problem of sustained resonance*, SIAM J. Appl. Math., 35 (1978), pp. 738–754.

[10] V. K. Melnikov, *On the stability of the center for time periodic perturbations*, Trans Moscow Math. Soc., 12 (1963), pp. 1–57.

[11] J. Murdock, *Some mathematical aspects of spin/orbit resonance*, Celestial Mech., 18 (1978), pp. 237–253.

[12] L. M. PERKO, *Higher order averaging and related methods for perturbed periodic and quasi-periodic systems*, SIAM J. Appl. Math., 17 (1968), pp. 698–723.
[13] C. ROBINSON AND J. MURDOCK, *Some mathematical aspects of spin/orbit resonance* II, Celestial Mech., 24 (1981), pp. 83–107.
[14] J. SANDERS, *On the passage through resonance*, this Journal, 10 (1979), pp. 1220–1243.
[15] _____, *Melnikov's method and averaging*, Celestial Mech., 28 (1982), pp. 171–181.
[16] W. KATH, *Necessay conditions for sustained reentry roll resonance*, SIAM J. Appl. Math., 43 (1983), pp. 314–324.
[17] R. HABERMAN, *Energy bounds for the slow capture by a center in sustained resonance*, SIAM J. Appl. Math., 43 (1983), pp. 244–256.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0002 $01.25/0

OSCILLATORY BIFURCATIONS IN SINGULAR PERTURBATION THEORY, I. SLOW OSCILLATIONS*

NEIL FENICHEL†

Abstract. We construct closed orbits for a singular perturbation problem near a nondegenerate equilibrium point of its reduced system, in case the linearization of the reduced system has a complex conjugate pair of pure imaginary eigenvalues. The mechanism for constructing the closed orbits is that of the familiar Hopf bifurcation problem, applied to a flow in a center manifold. The main content of this paper is an explicit computation, using a center manifold, of the parameters of the Hopf bifurcation in the singular case. A companion paper [SIAM J. Math. Anal., 14 (1983), pp. 868–874] studies a singular bifurcation problem in which closed orbits are constructed by a mechanism which is not related to a Hopf bifurcation.

Introduction. We study the structure of a singular perturbation problem near an equilibrium point of its reduced system. This structure is determined primarily by two sets of eigenvalues, the eigenvalues of the linearization of the reduced system and the eigenvalues of the fast flow.

In this paper we study the case in which the linearization of the reduced system has a pair of simple pure imaginary eigenvalues and the eigenvalues of the fast flow all have nonzero real parts. The orbit structure is close to the orbit structure of the regular Hopf bifurcation problem, so our main interest is in explicit computation of parameters.

In a companion paper [10] (this issue, pp. 868–874) we study the case in which the fast flow has a pair of simple pure imaginary eigenvalues. This case exhibits a richer structure than the previous case, because there is significant interaction between the fast flow and the reduced flow. We find oscillations, but the mechanism leading to oscillations is different from the mechanism of the regular Hopf bifurcation problem.

Oscillatory bifurcations appear in applications of singular perturbation theory. Examples are found in the formal computations of Poore [8] and Matkowsky and Sivashinsky [6], [7] on chemical reactors, in numerical computations of Feroe [3] on nerve conduction equations, and in Hastings' work on the Fitzhugh–Nagumo equations [4].

1. Statement of the problem. We study a C^{r+1} family of differential equations

$$z' = F(z,\varepsilon), \tag{1.1}$$

where $z \in R^{m+n}$, ε is a small real parameter, $2 \le r < \infty$, and $'$ denotes $\frac{d}{d\tau}$. We assume that (1.1) is singular for $\varepsilon = 0$ in the sense that $F(z,0)$ vanishes identically on an m-dimensional manifold $\mathcal{C}$. In particular, if (1.1) takes the special form

$$x' = \varepsilon f^R(x,y,\varepsilon), \qquad y' = g(x,y,\varepsilon), \tag{1.2}$$

where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, then for $\varepsilon = 0$ the right-hand side vanishes on an m-dimensional manifold near any point where g vanishes and g_y is nonsingular.

The linearization of (1.1) along a solution curve $z(t)$ is

$$\delta z' = F_z(z(t),\varepsilon)\delta z. \tag{1.3}$$

*Received by the editors January 26, 1982.

†Department of Mathematics, University of British Columbia, Vancouver, British Columbia V6T 1W5, Canada. This research was supported in part by the U. S. Army Research Office, and by the Natural Sciences and Engineering Research Council of Canada.

For $\varepsilon=0$ any point $z^0\in\mathfrak{S}$ is a constant solution of (1.1). The linearization (1.3) around such a constant solution is a constant coefficient equation

$$\delta z'=F_z(z^0,0)\delta z \tag{1.4}$$

with matrix $F_z(z^0,0)$. Because F vanishes identically on $\mathfrak{S}$, zero is an eigenvalue of $F_z(z^0,0)$ of multiplicity at least m. We define $\mathfrak{S}^R$ as the set of points $z^0\in\mathfrak{S}$ such that zero is an eigenvalue of $F_z(z^0,0)$ of multiplicity precisely m. For $z^0\in\mathfrak{S}^R$ we call the n nonzero eigenvalues the eigenvalues of the fast flow. In this paper we study (1.1) only near points of $\mathfrak{S}^R$.

The reduced system is a differential equation on $\mathfrak{S}$ derived by expanding F to first order in ε, projecting onto the tangent space of $\mathfrak{S}$, and rescaling time. In case all the eigenvalues of the fast flow have negative real parts we justify projecting onto the tangent space of $\mathfrak{S}$ because components along the complementary invariant subspace decay fast in the rescaled time.

The reduced system of (1.2) is obtained simply by setting $\varepsilon=0$ and solving for y in terms of x:

$$\dot{x}=f^R(x,y,0),\qquad 0=g(x,y,0), \tag{1.5}$$

where $\dot{}$ denotes $\frac{d}{dt}$, with

$$t=\varepsilon\tau. \tag{1.6}$$

In §2 we recall from [2] the computation of the reduced system.

We are interested in the orbit structure of (1.1) near an equilibrium point of the reduced system. At an equilibrium point of the reduced system, (1.1) has two well-defined sets of eigenvalues, the m eigenvalues of the linearization of the reduced system and the n eigenvalues of the fast flow. We say that (1.1) exhibits a slow oscillatory bifurcation if the linearization of the reduced system has a complex conjugate pair of pure imaginary eigenvalues and the eigenvalues of the fast flow all have nonzero real parts. We say that (1.1) exhibits a fast oscillatory bifurcation if it has a complex conjugate pair of pure imaginary eigenvalues of the fast flow.

In this paper we study only the slow oscillatory bifurcation. We compute a coefficient α in terms of the Taylor series of F, such that if α is nonzero and if some nonresonance conditions are satisfied, then (1.1) has a family of periodic solutions. This is a singular version of the Hopf bifurcation theorem.

We prefer to study (1.1), rather than (1.2), for two reasons. First, as we have argued in [2], (1.1) is more natural than (1.2). Second, for the study of fast oscillatory bifurcations, (1.2) is so restrictive that it does not exhibit the phenomena generically exhibited by (1.1). In §4 we restrict our attention to (1.2) in order to simplify a Taylor series computation.

2. The reduced system. To study (1.1) near a point in $\mathfrak{S}$ we choose coordinates $z=(x,y)\in\mathbb{R}^m\times\mathbb{R}^n$ in which $\mathfrak{S}$ is the graph of a C^{r+1} function $y=u^0(x)$. Then (1.1) takes the form

$$x'=f(x,y,\varepsilon),\qquad y'=g(x,y,\varepsilon), \tag{2.1}$$

subject to

$$f(x,u^0(x),0)=0,\qquad g(x,u^0(x),0)=0. \tag{2.2}$$

A computation in [2] shows that the reduced system on $\mathfrak{S}$ is

$$\dot{x}=f_\varepsilon+f_yh^{-1}u_x^0f_\varepsilon-f_yh^{-1}g_\varepsilon, \tag{2.3a}$$

$$\dot{y}=u_x\big(f_\varepsilon+f_yh^{-1}u_x^0f_\varepsilon-f_yh^{-1}g_\varepsilon\big), \tag{2.3b}$$

where

$$h=g_y-u_x^0 f_y, \tag{2.4}$$

$$u_x^0=-g_y^{-1}g_x, \tag{2.5}$$

and $\dot{}$ denotes $\frac{d}{dt}$. A further computation shows that the eigenvalues of the fast flow are precisely the eigenvalues of h.

We may take x as a coordinate on $\mathfrak{S}$. Then (2.3a) governs the flow of the reduced system and (2.3b) expresses the invariance of $\mathfrak{S}$.

We now assume that the origin in $\mathbb{R}^m\times\mathbb{R}^n$ is an equilibrium point of the reduced system. The linearization of (2.1) at the origin for $\varepsilon=0$ has an m-dimensional null space and an n-dimensional complementary invariant subspace. We assume that these are tangent to the x-axis and the y-axis, respectively, so that

$$f_x(0,0,0)=0, \tag{2.6}$$

$$f_y(0,0,0)=0 \tag{2.7}$$

and

$$g_x(0,0,0)=0. \tag{2.8}$$

Then the linearization of the reduced system (2.3a) at the origin is

$$\delta\dot{x}=k\,\delta x, \tag{2.9}$$

with

$$k=f_{\varepsilon x}(0,0,0)-f_{yx}(0,0,0)h^{-1}(0)g_\varepsilon(0,0,0), \tag{2.10}$$

and

$$h(0)=g_y(0,0,0). \tag{2.11}$$

We assume:

(H1) For some real nonzero ω, $i\omega$ and $-i\omega$ are simple eigenvalues of the linearization of the reduced equation at the origin. No other eigenvalue of the linearization of the reduced equation at the origin is an integer multiple of $i\omega$. In particular, zero is not an eigenvalue of k, so k is invertible and the origin is an isolated equilibrium point of the reduced equation.

3. Center manifolds. We now assume:

(H2) The eigenvalues of the fast flow at the origin all have nonzero real parts. That is, the eigenvalues of $g_y(0,0,0)$ all have nonzero real parts. This is a hyperbolicity assumption for (1.1) normal to $\mathfrak{S}$.

The hypothesis (H2) suggests that we should use a center manifold to reduce the singular problem to a regular problem. We extend (2.1) to

$$x'=f(x,y,\varepsilon),\quad y'=g(x,y,\varepsilon),\quad \varepsilon'=0, \tag{3.1}$$

where the equation $\varepsilon'=0$ reflects the role of ε as a parameter.

The linearization of (3.1) at the origin is

$$\begin{pmatrix}\delta_x\\ \delta_y\\ \delta_\varepsilon\end{pmatrix}'=\begin{bmatrix}0 & 0 & f_\varepsilon\\ 0 & g_y & g_\varepsilon\\ 0 & 0 & 0\end{bmatrix}\begin{pmatrix}\delta_x\\ \delta_y\\ \delta_\varepsilon\end{pmatrix}. \tag{3.2}$$

Zero is an eigenvalue of algebraic multiplicity $m+1$ for the matrix of (3.1); the remaining n eigenvalues have nonzero real parts by (H2).

The center manifold theorem asserts that (3.2) has a C^r invariant manifold $\mathcal{C}$, the graph of a function

$$y=u(x,\varepsilon), \tag{3.3}$$

with

$$u(x,0)=u^0(x). \tag{3.4}$$

Center manifolds in the context of singular perturbation theory are discussed in [5] and [2]; a center manifold theorem which includes the existence of $\mathcal{C}$ is proved in [2].

On $\mathcal{C}$, the evolution of (2.1) satisfies

$$x'=f(x,u(x,\varepsilon),\varepsilon), \quad u(x,\varepsilon)'=g(x,u(x,\varepsilon),\varepsilon), \quad \varepsilon'=0. \tag{3.5}$$

We take (x,ε) as coordinates in $\mathcal{C}$. Because ε is constant, the evolution of (3.1) on $\mathcal{C}$ is governed by

$$x'=f(x,u(x,\varepsilon),\varepsilon). \tag{3.6}$$

The second equation of (3.5) gives the invariance condition

$$u_x(x,\varepsilon)f(x,u(x,\varepsilon),\varepsilon)=g(x,u(x,\varepsilon),\varepsilon). \tag{3.7}$$

The function u has a unique asymptotic expansion at $\varepsilon=0$. See Wan [9]. We use (3.7) to compute the asymptotic expansion of u.

Note that

$$f(x,u(x,0),0)\equiv 0, \tag{3.8}$$

so

$$f(x,u(x,\varepsilon),\varepsilon)=O(\varepsilon). \tag{3.9}$$

Hence the rescaled equation

$$\dot{x}=\varepsilon^{-1}f(x,u(x,\varepsilon),\varepsilon) \tag{3.10}$$

is C^r even at $\varepsilon=0$. It follows from the general theory in [2] that (3.10) is a C^r perturbation of the reduced equation (2.3a). To see this directly, note that by (3.8)

$$\begin{aligned}\varepsilon^{-1}f(x,u(x,\varepsilon),\varepsilon)&=\varepsilon^{-1}[f(x,u(x,\varepsilon),\varepsilon)-f(x,u(x,0),0)]\\ &\to f_y(x,u(x,0),0)u_\varepsilon(x,0)+f_\varepsilon(x,u(x,0),0)\end{aligned} \tag{3.11}$$

as $\varepsilon\to 0$. It follows by differentiating (3.7) at $\varepsilon=0$ that

$$u_\varepsilon=-h^{-1}g_\varepsilon+h^{-1}u_x f_\varepsilon \tag{3.12}$$

and

$$u_x=-g_y^{-1}g_x, \tag{3.13}$$

so (3.10) tends to (2.3a) as $\varepsilon\to 0$.

Because (3.10) is a regular perturbation of (2.3a), it follows from (H1) that the equilibrium point at the origin perturbs smoothly to an equilibrium point $\xi(\varepsilon)$ of (3.10). We have

$$f(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)\equiv 0. \tag{3.14}$$

The linearization of (3.10) at $x(\varepsilon)$ is

$$\delta\dot{x}=\varepsilon^{-1}L(\varepsilon)\delta x, \tag{3.15}$$

where

$$L(\varepsilon)=f_x(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)+f_y(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)u_x(\xi(\varepsilon),\varepsilon). \tag{3.16}$$

By (2.2) and (3.4), $L(0)=0$. Because (3.10) is a C^r perturbation of (2.3a), $\varepsilon^{-1}L(\varepsilon)\to k$ as $\varepsilon\to 0$. Let $\lambda(\varepsilon)$ be the eigenvalue of $\varepsilon^{-1}L(\varepsilon)$ which continues the eigenvalue $i\omega$ of k. The coefficient α mentioned in §1 is $\mathrm{Re}\lambda'(0)$. Our aim is to compute α in terms of derivatives of f and g at the origin. Let

$$l=\frac{d}{d\varepsilon}\left(\varepsilon^{-1}L(\varepsilon)\right)\Big|_{\varepsilon=0}. \tag{3.17}$$

The main step in the computation of α is the computation of l. From the Taylor series for L we see that

$$l=\tfrac{1}{2}L''(0). \tag{3.18}$$

Assume now that we have chosen a basis for $\mathbb{R}^m$ in which k has the block diagonal form

$$k=\begin{bmatrix} A & 0 \\ 0 & \omega J \end{bmatrix} \tag{3.19}$$

where

$$J=\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{3.20}$$

Write $A=(A_{ij})$ and $l=(l_{ij})$. Let

$$p(\lambda,\varepsilon)=\det\left[\lambda I-\varepsilon^{-1}L(\varepsilon)\right]. \tag{3.21}$$

Then

$$p(\lambda,\varepsilon)=p^1(\lambda,\varepsilon)p^2(\lambda,\varepsilon)+O(\varepsilon^2), \tag{3.22}$$

where

$$p^1(\lambda,\varepsilon)=\det(\lambda I-A)+O(\varepsilon) \tag{3.23}$$

and

$$p^2(\lambda,\varepsilon)=\det\left(\lambda I-\omega J-\varepsilon\begin{bmatrix} l_{n-1,n-1} & l_{n-1,n} \\ l_{n,n-1} & l_{n,n} \end{bmatrix}\right). \tag{3.24}$$

We know that

$$p(\lambda(\varepsilon),\varepsilon)=0, \tag{3.25}$$

$$\lambda(0)=i\omega, \tag{3.26}$$

$$p^1(i\omega,0)\neq 0 \tag{3.27}$$

and

$$p^2(i\omega,0)=0. \tag{3.28}$$

From (3.25) we have

$$p_\lambda(i\omega,0)\lambda'(0)+p_\varepsilon(i\omega,0)=0, \tag{3.29}$$

so by (3.21)

$$\begin{aligned}\lambda'(0)&=-p_\varepsilon^2(i\omega,0)/p_\lambda^2(i\omega,0)\\ &=\tfrac{1}{2}(l_{n-1,n-1}+l_{n,n})-\tfrac{1}{2}i(l_{n-1,n}-l_{n,n-1}).\end{aligned} \tag{3.30}$$

Hence

$$\alpha=\tfrac{1}{2}(l_{n-1,n-1}+l_{n,n}). \tag{3.31}$$

We now state a singular version of the Hopf bifurcation theorem.

THEOREM. *Let* (2.1) *satisfy* (H1) *and* (H2). *Suppose* $\alpha\neq 0$. *Then there is a smooth* 2*-dimensional manifold* $\mathfrak{M}$ *in* $\mathbb{R}^{m+n}\times\mathbb{R}$ *which is invariant under* (3.1). $\mathfrak{M}$ *contains the origin; all other points in* $\mathfrak{M}$ *lie either on periodic orbits of* (3.1) *with* $\varepsilon\neq 0$ *or on periodic orbits of the reduced system with* $\varepsilon=0$.

Proof. This theorem follows directly from the Hopf bifurcation theorem of Crandall and Rabinowitz [1] applied to (3.10).

4. Computation of l. To make the results of the previous section useful we must compute l in terms of the derivatives of f and g at the origin. From (3.16) and (3.18) we express l in terms of derivatives of $f, u,$ and ξ at the origin. From (3.14) we compute derivatives of ξ in terms of derivatives f and u, and from (3.7) we compute derivatives of u in terms of derivatives of f and g.

To exhibit this computation in a reasonable space we restrict ourselves now to equations in the special form (1.2). Then (3.7), (3.14), and (3.16) take the forms

$$g(x,u(x,\varepsilon),\varepsilon)=\varepsilon u_x(x,\varepsilon)f^R(x,u(x,\varepsilon),\varepsilon), \tag{4.1}$$

$$f^R(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)=0, \tag{4.2}$$

$$\varepsilon^{-1}L(\varepsilon)=f_x^R(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)+f_y^R(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)u_x(\xi(\varepsilon),\varepsilon). \tag{4.3}$$

From (3.16) we have the simplified expression

$$\begin{aligned} l=&f_{xx}^R\xi_\varepsilon+f_{xy}^R(u_x\xi_\varepsilon+u_\varepsilon)+f_{x\varepsilon}^R\\ &+\left[f_{yx}^R\xi_\varepsilon+f_{yy}^R(u_x\xi_\varepsilon+u_\varepsilon)+f_{y\varepsilon}^R\right]u_x+f_y^R(u_{xx}\xi_\varepsilon+u_{x\varepsilon}), \end{aligned} \tag{4.4}$$

where all derivatives are evaluated at the origin. We know from (2.8) and from (2.5) and (3.4) that g_x and u_x vanish at the origin. Hence (4.4) simplifies to

$$l=f_{xx}^R\xi_\varepsilon+f_{xy}^Ru_\varepsilon+f_{x\varepsilon}^R+f_y^R(u_{xx}\xi_\varepsilon+u_{x\varepsilon}). \tag{4.5}$$

From (3.14) we have

$$f^R(\xi(\varepsilon),u(\xi(\varepsilon),\varepsilon),\varepsilon)=0, \tag{4.6}$$

so

$$f_x^R\xi_\varepsilon+f_y^R(u_x\xi_\varepsilon+u_\varepsilon)+f_\varepsilon^R=0. \tag{4.7}$$

Note that

$$k=f_x^R(0,0,0) \tag{4.8}$$

is invertible by (H1), so we have

$$\xi_\varepsilon=-\left(f_x^R\right)^{-1}\left(f_y^Ru_\varepsilon+f_\varepsilon^R\right). \tag{4.9}$$

Differentiating (4.1) with respect to ε at $\varepsilon=0$ yields

$$g_yu_\varepsilon+g_\varepsilon=0, \tag{4.10}$$

so

$$u_\varepsilon=-g_y^{-1}g_\varepsilon \tag{4.11}$$

and

$$\xi_\varepsilon = -\left(f_x^R\right)^{-1}\left(-f_y^R g_y^{-1} g_\varepsilon + f_\varepsilon^R\right). \tag{4.12}$$

Differentiating (4.1) with respect to x yields

$$g_x + g_y u_x = \varepsilon u_{xx} f^R + \varepsilon u_x\left(f_x^R + f_y^R u_x\right). \tag{4.13}$$

At $\varepsilon = 0$ we have

$$u_x = -g_y^{-1} g_x = 0. \tag{4.14}$$

Differentiating (4.1) with respect to x and ε at $\varepsilon = 0$ yields

$$g_{xx} + g_y u_{xx} = 0 \tag{4.15}$$

and

$$g_{xy} u_\varepsilon + g_{x\varepsilon} + g_y u_{x\varepsilon} = 0. \tag{4.16}$$

Note that f^R vanishes at the origin, because the origin is an equilibrium point of the reduced equation. From (4.11), (4.15) and (4.16) we have

$$u_{xx} = -g_y^{-1} g_{xx} \tag{4.17}$$

and

$$u_{x\varepsilon} = -g_y^{-1}\left(g_{xy} u_\varepsilon + g_{x\varepsilon}\right) = -g_y^{-1}\left(-g_{xy} g_y^{-1} g_\varepsilon + g_{x\varepsilon}\right). \tag{4.18}$$

Finally, we have

$$\begin{aligned} l = {} & -f_{xx}^R\left(f_x^R\right)^{-1}\left(-f_y^R g_y^{-1} g_\varepsilon + f_\varepsilon^R\right) - f_{xy}^R g_y^{-1} g_\varepsilon + f_{x\varepsilon}^R \\ & + f_y^R\left[g_y^{-1} g_{xx}\left(f_x^R\right)^{-1}\left(-f_y^R g_y^{-1} g_\varepsilon + f_\varepsilon^R\right) - g_y^{-1}\left(-g_{xy} g_y^{-1} g_\varepsilon + g_{x\varepsilon}\right)\right]. \end{aligned} \tag{4.19}$$

This completes the computation of l for equations in the special form (1.2). We are content to leave the computation in the general case to the reader.

Acknowledgments. I wish to thank Jack Hale, Dave Schaeffer and Aubrey Poore for discussions leading to this paper.

REFERENCES

[1] M. CRANDALL AND P. RABINOWITZ, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rational Mech. Anal., 67 (1977), pp. 53–72.

[2] N. FENICHEL, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[3] J. FEROE, *Existence and stability of multiple impulse solutions of a nerve equation*, SIAM J. Applied Math., 42 (1982), pp. 235–246.

[4] S. HASTINGS, *Single and multiple pulse waves for the Fitzhugh–Nagumo equations*, SIAM J. Applied Math., 42 (1982), pp. 247–260.

[5] N. KOPELL, *Waves, shocks and target patterns in an oscillating chemical reagent*, Proc. 1976 Symposium of Non-Linear Diffusion Equations at Houston, Pitman Press, London, 1977.

[6] B. MATKOWSKY AND G. SIVASHINSKY, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Applied Math., 35 (1978), pp. 465–478.

[7] _____, *On oscillatory necking in polymers*, Appl. Math. Tech. Rep. 7705, Northwestern University, Evanston, IL, 1977.

[8] A. POORE, Personal communication.

[9] Y.-H. WAN, *On the uniqueness of invariant manifolds*, J. Differential Equations, 24 (1977), pp. 268–273.

[10] N. FENICHEL, *Oscillatory bifurcations in singular perturbation theory*, II. *Fast oscillations*, this Journal, this issue, pp. 868–874.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0003 $01.25/0

OSCILLATORY BIFURCATIONS IN SINGULAR PERTURBATION THEORY, II. FAST OSCILLATIONS*

NEIL FENICHEL†

Abstract. We construct closed orbits for a singular perturbation problem near a nondegenerate equilibrium point of its reduced system, in case the linearization of the fast part of the flow has a complex conjugate pair of pure imaginary eigenvalues. The mechanism for constructing the closed orbits is essentially different from the mechanism of a Hopf bifurcation.

Introduction. We study the orbit structure of a singular perturbation problem near a nondegenerate equilibrium point of its reduced system. At the equilibrium point there are two well defined sets of eigenvalues; these correspond to the linearization of the reduced system and to the linearization of the fast flow. A reasonably complete analysis of the orbit structure is straightforward if all eigenvalues in both sets have nonzero real parts; see [2, Thm. 12.2]. The flow in this case is closely related to the flow of a regular system near a hyperbolic equilibrium point. If some of the eigenvalues have real parts equal to zero, however, a singular bifurcation is expected. This bifurcation may be classified as a slow bifurcation, in case the eigenvalues with zero real parts correspond to the linearization of the reduced system, or as a fast bifurcation, in case the eigenvalues with zero real parts correspond to the linearization of the fast flow.

In an earlier paper [1] (this issue, pp. 861–867) we studied the slow bifurcation which occurs when the linearization of the reduced system has a complex conjugate pair of eigenvalues on the imaginary axis. In that case the orbit structure is reduced to the structure of an ordinary Hopf bifurcation, so the main interest is in computation of the terms which control the structure of the Hopf bifurcation. In the present paper we study the more difficult case in which the linearization of the fast flow has a complex conjugate pair of pure imaginary eigenvalues. We compute a matrix and a number in terms of the Taylor series at the equilibrium point, such that periodic orbits bifurcate from the equilibrium point if the matrix is nonsingular and the number is positive. The bifurcation is not a Hopf bifurcation, however, because the mechanism for constructing the periodic orbits is essentially different from the mechanism of a Hopf bifurcation.

The entire analysis presented below depends upon the geometric singular perturbation theory developed in [2]. The conditions we require are natural in the geometric theory, but never are satisfied in the standard theory. In contrast, slow bifurcation phenomena always occur in center manifolds and hence are identical in the geometric theory and in the standard theory. For slow bifurcations in which the fast flow is stable [2, Thm. 11.1] even guarantees that the geometric theory and the standard theory agree up to a smooth change of coordinates.

Fast bifurcation problems are especially rich because they combine the complexities of bifurcation theory and of singular pertubation theory. The exchange of stability problem studied by Lebovitz and Schaar [4] is another example. In these problems the normal forms constructed in [2] show that the standard theory omits certain terms which are insignificant in case the fast flow is stable, but which are essential in the study of fast bifurcations. We hope this study will focus attention on the differences

*Received by the editors January 26, 1982.

†Department of Mathematics, University of British Columbia, Vancouver, British Columbia, V6T 1W5, Canada. This research was supported in part by the U.S. Army Research Office and by the Natural Sciences and Engineering Research Council of Canada.

between geometric singular perturbation theory and the standard theory, and so clarify the role of geometry in this field.

1. A singular bifurcation problem. We recall the standard form of a singular perturbation problem, and the geometric form developed in [2]. In the standard form, the variables are separated into two vectors, x in R^m, and y in R^n, satisfying a system of differential equations of the form

$$x' = f(x,y,\varepsilon), \qquad \varepsilon y' = g(x,y,\varepsilon), \tag{1}$$

where ε is a small real parameter, and $'$ stands for $\frac{d}{dt}$. In order to understand the geometry of (1) it is convenient to introduce the rescaled time T, defined by $t=\varepsilon T$. Then (1) takes the form

$$x' = \varepsilon f(x,y,\varepsilon), \qquad y' = g(x,y,\varepsilon), \tag{2}$$

where now $'$ denotes $\frac{d}{dT}$. Let

$$E = \{(x,y): g(x,y,0)=0\}.$$

For $\varepsilon=0$ the first equation of (1) reduces to $g(x,y,0)=0$, so it is natural to study (1) in a neighborhood of E.

E consists entirely of equilibrium points for the rescaled system (2). If the Jacobian $\frac{\partial g}{\partial y}$ is nonsingular on E, then E is a manifold of dimension m, the graph of a smooth function $y=u^0(x)$. The singularity of the system (1) is reflected in the fact that the equilibrium points of the rescaled system (2) are not isolated, but instead form a smooth manifold.

The situation just described is the starting point of the geometric approach to singular perturbation theory. Following [2], we define a singular perturbation problem to be a family of differential equations depending on a small parameter ε, with the property that for $\varepsilon=0$ the system has a manifold E of dimension $m>0$ consisting entirely of equilibrium points. Thus we study a system

$$w' = h(w,\varepsilon) \tag{3a}$$

with

$$h(w,0)=0 \quad \text{for all } w \text{ in } E. \tag{3b}$$

The system (2) above has the form (3), with $w=(x,y)$, $h=(\varepsilon f, g)$, and E the zero set of g.

From (3b) it follows that the derivative of h along any direction tangent to E is zero. Thus 0 is an eigenvalue of the Jacobian $\frac{\partial h}{\partial w}$, with multiplicity at least equal to m, the dimension of E. For many applications the natural stability condition is to require that the remaining eigenvalues lie in the left plane. (This corresponds to the usual requirement that the eigenvalues of $\frac{\partial g}{\partial y}$ lie in the left half plane.) Under this stability hypothesis there is a smooth change of coordinates transforming (3) into the form (2); see [2, Thm. 11.1]. The eigenvalues of $\frac{\partial g}{\partial y}$ then lie in the left half plane.

The subject of the present paper is the structure of (3) near a point at which the Jacobian $\frac{\partial h}{\partial w}$ has zero as an eigenvalue of multiplicity exactly m, but also has one complex conjugate pair of nonzero eigenvalues on the imaginary axis. This means that all neutral behavior in (3) comes from translation along E, but that there is a two-dimensional plane of fast rotation transversal to E. If one nondegeneracy condition is satisfied, and if a certain coefficient is positive, we show that (3) has a family of periodic orbits. These periodic orbits have radius of order $|\varepsilon|$ and are formed by a balance between the evolution of the slow variables and the evolution of the the radial

part of fast variables. In a Hopf bifurcation the periodic orbits have radius of order $|\varepsilon|^{1/2}$ and are formed by a balance between weak linear terms and weak nonlinear terms. Thus the mechanism we are studying for the construction of periodic orbits is fundamentally different from the mechanism of a Hopf bifurcation.

The conditions we impose in order to construct periodic orbits never are satisfied by systems of the form (2). This means at least that the derivation of (2) should be verified any time (2) is applied to a singular perturbation problem which exhibits bifurcation in the fast variables. It further suggests that (2) may be inappropriate for the study of such problems.

2. Normal forms. Consider a singular perturbation problem of the form (3), in a neighborhood of a point p in E. Assume that zero is an eigenvalue of multiplicity exactly m for the Jacobian $\frac{\partial h}{\partial w}(p)$, and assume that all other eigenvalues of $\frac{\delta h}{\delta w}(p)$ have nonzero real parts, except for one pair $\pm ik^0$ of simple pure imaginary eigenvalues. By [2, Thm. 11.1], in a neighborhood of p in R^{m+n} there are coordinates x in R^m and $y=(y_1,y_2,y_3)$ in R^n such that (3) takes the form

$$x'=f(x,y_1,y_2,y_3,\varepsilon), \qquad y'=g(x,y_1,y_2,y_3,\varepsilon) \tag{4}$$

with functions f and $g=(g_1,g_2,g_3)$ satisfying

$$f(x,0,0,0,0)=0, \qquad g(x,0,0,0,0)=0 \tag{5}$$

and

$$\begin{aligned} &g_1(x,0,y_2y_3,\varepsilon)=0, && g_3(x,y_1,y_2,0,\varepsilon)=0,\\ &D_2f(x,y_1,y_2,0,\varepsilon)=0, && D_2g(x,y_1,y_2,0,\varepsilon)=0,\\ &D_3f(x,0,0,0,0,0)=0 && \\ &D_4f(x,0,y_2,y_3,\varepsilon)=0, && D_4g(x,0,y_2,y_3,\varepsilon)=0, \end{aligned} \tag{6}$$

and

$$\frac{\partial g}{\partial y}(x,0,0,0,0)=\begin{bmatrix} A_1(x) & 0 & 0\\ 0 & A_2(x) & 0\\ 0 & 0 & A_3(x)\end{bmatrix}, \tag{7}$$

where the eigenvalues of $A_1(0)$ lie in the right half plane, the eigenvalues of $A_2(0)$ lie on the imaginary axis and the eigenvalues of $A_3(0)$ lie in the left half plane. Note that y_2 is in R^2, and the eigenvalues of $A_2(0)$ are precisely $\pm ik^0$.

The first two equations of (6) show that the points $(x,0,y_2,0,0)$ form an invariant set for the flow of (4). This set is a center manifold C of dimension $m+2$, and by [2, Thm. 9.1] the center manifold contains all invariant sets near p. Stability of the invariant sets is related to stability within C, as detailed in [2, Thm. 9.1]. Furthermore, the center manifold has a unique asymptotic expansion, so formal series computations are easily justified. Hence to study the periodic orbits of (4) we are free to restrict attention to C. We now make this restriction by assuming that $n=2$, and we denote y_2, g_2 and A_2 simply as y, g and A. Then (4) becomes

$$x'=f(x,y,\varepsilon), \qquad y'=g(x,y,\varepsilon) \tag{8}$$

subject to

$$\begin{aligned} &f(x,0,0)=0, && g(x,0,0)=0,\\ &D_2f(x,0,0)=0, && D_2g(x,0,0)=A(x). \end{aligned} \tag{9}$$

Because the eigenvalues $\pm ik^0$ are simple, the eigenvalues $\lambda(x)$ and $\bar{\lambda}(x)$ of $A(x)$ vary smoothly with x. The corresponding eigenvectors, suitably normalized, also vary smoothly with x, so by a smooth x-dependent linear change of the y coordinates we may assume that the matrix $A(x)$ takes the form

$$\begin{bmatrix} a(x) & -k(x) \\ k(x) & a(x) \end{bmatrix}.$$

We replace the real 2-vector y by one complex variable z whose real and imaginary parts are the first and second components, respectively, of y. Then we expand functions of y as series in z and its complex conjugate $\bar{z}$. This puts (8) in the form

$$x' = f(x,z,\bar{z},\varepsilon), \qquad z' = g(x,z,\bar{z},\varepsilon) \tag{10}$$

subject to

$$\begin{aligned} &f(x,0,0,0)=0, && g(x,0,0,0)=0, \\ &D_2 f(x,0,0,0)=0, && D_3 f(x,0,0,0)=0, \\ &D_2 g(x,0,0,0)=\lambda(x), && D_3 g(x,0,0,0)=0. \end{aligned} \tag{11}$$

3. Averaging. We now simplify (10) using an algebraic averaging procedure as in Segrè [5] and Sacker [6]. This procedure also has been used in computations of Hassard and Wan [3]. At each step our computations are formal; they can be made rigorous by restricting to a small neighborhood of the origin and appealing to the implicit function theorem.

For $\varepsilon = 0$, the Taylor series of $f(x,z,\bar{z},0)$ has no constant term, no terms in x alone and no linear term in z or $\bar{z}$. To lowest order f has the form

$$f(x,z,\bar{z},0) = a_1 xz + a_2 x\bar{z} + b_1 z^2 + b_2 z\bar{z} + b_3 \bar{z}^2 + \cdots.$$

To simplify f we try a formal substitution:

$$x_1 = x + r_1 xz + r_2 x\bar{z} + s_1 z^2 + s_2 z\bar{z} + s_3 \bar{z}^2 + \cdots.$$

This yields

$$\begin{aligned} x_1' &= x' + r_1 x'z + r_1 xz' + r_2 x'\bar{z} + r_2 x\bar{z}' \\ &\quad + 2s_1 zz' + s_2 z'\bar{z} + s_2 z\bar{z}' + 2s_3 \bar{z}^{2\prime} + \cdots \\ &= a_1 xz + a_2 x\bar{z} + b_1 z^2 + b_2 z\bar{z} + b_3 \bar{z}^2 \\ &\quad + ik^0 r_1 xz - ik^0 r_2 x\bar{z} + 2ik^0 s_1 z^2 - 2ik^0 s_3 \bar{z}^2 + \{\text{higher order terms}\} \\ &= (a_1 + ik^0 r_1)xz + (b_1 - ik^0 r_1)x\bar{z} + (b_1 + 2ik^0 s_1)z^2 + b_2 z\bar{z} \\ &\quad + (b_3 - 2ik^0 s_3)\bar{z}^2 + \{\text{higher order terms}\}. \end{aligned}$$

We select r_1, r_2, s_1, and s_3 to make the coefficients of xz, $x\bar{z}$, z^2, and $\bar{z}^2$ vanish. Note that the coefficient of $z\bar{z}$ is invariant. This is natural, as $z\bar{z} = |z|^2$ is varying slowly, while z^2 and $\bar{z}^2$ oscillate rapidly and hence on average their effects are small.

A significant difference between geometric singular perturbation theory and the usual singular perturbation theory appears in this simple computation. The invariant b_2 always is zero for a system of the form (2), and generically is nonzero for a system of the form (3). Setting b_2 equal to zero seems unnatural when studying a fast bifurcation, as it means ignoring the effect on the slow variable x of the radial drift of the fast variable z.

We now rename x_1 as x and try a similar coordinate transformation in order to simplify z' for $\varepsilon=0$. Note that all terms in z' which are linear in z appear already in $\lambda(x)z$, and no terms linear in $\bar{z}$ appear. The first terms we try to simplify are the terms in z^2, $z\bar{z}$, and $\bar{z}^2$. A computation like the one above shows that all of these can be removed by a transformation:

$$z_1=z+s_1z^2+s_2z\bar{z}+s_3\bar{z}^2.$$

We omit the computation. After solving for s_1, s_2, and s_3 to remove all the quadratic terms in z and $\bar{z}$, we relabel z_1 as z and expand $\lambda(x)$ as $ik^0+Lx+O_2(x)$. Then for $\varepsilon=0$ we have the system (10) in the form

$$\begin{aligned} x'&=b|z|^2+O_2(x,z,\bar{z})z+O_2(x,z,\bar{z})\bar{z},\\ z'&=ik^0z+Lxz+O_2(x,z,\bar{z})z+O_2(x,z,\bar{z})z. \end{aligned} \tag{12}$$

4. Construction of periodic orbits. From (12) and [2, Lemma 5.4], the reduced system of (10) satisfies the equation

$$x'=D_4f(x,0,0,0) \tag{13}$$

in the plane $z=0$. This means that the reduced system is just the first order expansion of f in ε, in the special coordinates in which (11) holds. By hypothesis, the origin in this coordinate system is a nondegenerate equilibrium point of the reduced system, so the linear term in ε in the Taylor series for f must vanish, and the coefficient of the quadratic term in εx must be invertible. This means that the expansion for (x',z') to quadratic terms in $(x,z,\bar{z},\varepsilon)$ has the form

$$\begin{aligned} x'&=b|z|^2+\varepsilon C_1x+\varepsilon C_2z+\varepsilon^2C_3+\cdots,\\ z'&=ik^0z+Lxz+\varepsilon c+\varepsilon B_1x+\varepsilon B_2z+\varepsilon B_3\bar{z}+\varepsilon^2B_4+\cdots. \end{aligned} \tag{14}$$

The matrix C_1 is just the linearization of the reduced system at the origin, so it is invertible and also invariant up to similarity. As above, we try to remove as many terms as possible by means of formal substitutions whose Taylor series begin with the identity. Replacing x by $x-\varepsilon C_1^{-1}C_3$ removes the term ε^2C_3, without otherwise changing the form of (14); we make this replacement. Because C_1 is invariant up to similarity we then try only to simplify x by a transformation

$$x_1=x+\varepsilon r_1z+\varepsilon r_2\bar{z}+\cdots.$$

This gives

$$x_1'=b|z|^2+\varepsilon C_1x+\varepsilon C_2z+\varepsilon C_3\bar{z}+\varepsilon ik^0r_2z-\varepsilon ik^0r_3\bar{z}+\cdots.$$

By choosing r_1 and r_2 we remove the terms in εz and $\varepsilon\bar{z}$.

To simplify z', we try a transformation of the form

$$z_1=z+\varepsilon r+\varepsilon s_1x+\varepsilon s_2z+\varepsilon s_3\bar{z}+\varepsilon^2s_4+\cdots.$$

Then

$$\begin{aligned} z_1'&=z'+\varepsilon s_1x'+\varepsilon s_2z'+\varepsilon s_3\bar{z}'+\cdots\\ &=(ik^0+Lx)(z_1-\varepsilon r-\varepsilon s_1x-\varepsilon s_2z-\varepsilon s_3\bar{z}-\varepsilon^2s_4)\\ &\quad+\varepsilon c+\varepsilon B_1x+\varepsilon B_2z+\varepsilon B_3\bar{z}+\varepsilon^2B_4+\varepsilon ik^0s_2z-\varepsilon ik^0s_3\bar{z}+\cdots\\ &=ik^0z_1+\varepsilon(c-ik^0r)+\varepsilon(B_1-ik^0s_1)x\\ &\quad+\varepsilon B_2z+\varepsilon(B_3-2ik^0s_3)\bar{z}+\varepsilon^2(B_4-ik^0s_4)+\cdots. \end{aligned}$$

We choose r, s_1, s_3 and s_4 to make the terms in $\varepsilon, \varepsilon x, \varepsilon\bar{z}$ and ε^2 all vanish, and we replace the term $\varepsilon B_2 z$ by $\varepsilon B_2 z_1$, with only a higher order error. Then we rename z_1 as z, and C_1 as C. This puts (10) in the form

$$x' = b|z|^2 + \varepsilon Cx + \cdots, \qquad z' = ik^0 z + Lz + \varepsilon cz + \cdots. \tag{15}$$

Finally, we scale both x and z by the factor ε, letting $x = \varepsilon X$, $z = \varepsilon Z$. Then (15) takes the form

$$X' = \varepsilon\big(CX + b|Z|^2\big) + O_2(\varepsilon), \qquad Z' = \big(ik^0 + \varepsilon(c + LX)\big)Z + O_2(\varepsilon). \tag{16}$$

Note that if x and z are scaled by $-\varepsilon$ instead of ε, the signs of the quadratic terms $b|Z|^2$ and LXZ are reversed.

Now we introduce polar coordinates $Z = R\exp(is)$ to separate the rapid angular variation of Z from the slow radial drift. Let $c = c_1 + ic_2$ and $L = L_1 + iL_2$. Then (16) is transformed to

$$\begin{aligned} X' &= \varepsilon(CX + bR^2) + O_2(\varepsilon), \\ R' &= \varepsilon(c_1 + L_1 X)R + O_2(\varepsilon), \\ s' &= k^0 + \varepsilon(c_2 + L_2 X)R + O_2(\varepsilon). \end{aligned} \tag{17}$$

We analyze (16) by computing the Poincaré map along the positive real z axis from $s = 0$ to $s = 2\pi$. Let $T(\varepsilon) = T(X, R, \varepsilon)$ denote the time for (17) to evolve from $(X, R, 0)$ to a point with $s = 2\pi$. From the equation for s' in (17) we see that $T(\varepsilon)$ can be expanded as

$$T(\varepsilon) = 2\pi/k^0 + O(\varepsilon).$$

Hence the Poincaré map takes (X, R) to (X_1, R_1), where

$$\begin{aligned} X_1 &= X + (2\pi/k^0)\varepsilon(CX + bR^2) + O_2(\varepsilon), \\ R_1 &= R + (2\pi/k^0)\varepsilon(c_1 + L_1 X)R + O_2(\varepsilon). \end{aligned} \tag{18}$$

The Poincaré map has a nontrivial fixed point if the equations

$$CX + bR^2 = O(\varepsilon), \qquad L_1 X = -c_1 + O(\varepsilon) \tag{19}$$

have a nonzero solution with R positive. For this we require the nondegeneracy condition that the matrix

$$M = \begin{bmatrix} C & b \\ L_1 & 0 \end{bmatrix} \tag{20}$$

be invertible. This condition cannot be satisfied if b or L_1 is zero. In particular, it never is satisfied for a singular perturbation problem in the form (4).

Recall that C is invertible, and let $a = L_1 C^{-1} b$. then the inverse of the matrix M is given explicitly by

$$M^{-1} = \begin{bmatrix} C^{-1} - a^{-1}C^{-1}bLC^{-1} & a^{-1}C^{-1}b \\ a^{-1}Lc^{-1} & -a^{-1} \end{bmatrix}, \tag{21}$$

and we find that

$$R^2 = a^{-1}c_1 + O(\varepsilon),$$

with a similar expression for X. If $a^{-1}c_1$ is negative, there are no periodic orbits with x and z of order ε. If $a^{-1}c_1$ is positive, then for each small ε, positive or negative, there is

one periodic orbit with x and z of order ε. In this case the set of periodic orbits forms a double cone in (x, z, ε) space.

5. Stability. The previous computations lead to simple stability criteria. If the Poincaré map is expressed in terms of X and R^2, the first order terms in ε are

$$X_1 = X + \left(\frac{2\pi}{k^0}\right)\varepsilon(CX + bR^2) + O_2(\varepsilon), \tag{22}$$

$$R_1^2 = R^2 + 2\left(\frac{2\pi}{k^0}\right)\varepsilon(c_1 + L_1 X)R^2 + O_2(\varepsilon).$$

The matrix of the linearization of the Poincaré map around the fixed point (X_0, R_0^2) is given, to first order in ε, by the identity plus $(2\pi/k^0)\varepsilon$ times

$$M_1 = \begin{bmatrix} C & b \\ 2L_1 a^{-1} c_1 & 0 \end{bmatrix}. \tag{23}$$

Within the center manifold, the periodic orbits are stable if all eigenvalues of this matrix lie inside the unit circle, and unstable if some eigenvalues lie outside the unit circle. Within the larger space R^{m+n}, the stability of the periodic orbits is determined by the stability of periodic orbits within the center manifold and the stability of the center manifold in R^{m+n}. For each eigenvalue of the fast flow which lies in the right half plane, the periodic orbits pick up one unstable dimension.

6. Summary.

We summarize the results of the preceding computations in the following

THEOREM. *Let a singular perturbation problem of the form* (3) *be given, with an equilibrium manifold of dimension m. Let p be a nondegenerate equilibrium point of the reduced system of* (3). *Suppose* 0 *is an eigenvalue of mulitiplicity exactly m of the linearization of* (3) *at p, and that the linearization also has one complex conjugate pair of pure imaginary eigenvalues. Suppose the remaining eigenvalues of the linearization lie off the imaginary axis. Then there is a matrix M, and there are numbers a and* c_1, *such that if M is invertible and if* $a^{-1}c_1$ *is positive then there is a family of periodic orbits of* (3) *whose radii are of order* $|\varepsilon|$; *if M is invertible and if* $a^{-1}c_1$ *is negative there are no periodic orbits of* (3) *whose radii are of order* $|\varepsilon|$. *The stability of the periodic orbits is determined by the eigenvalues of the linearization of* (3) *at p which lie off the imaginary axis, and by the eigenvalues of a matrix* M_1. *The matrices M and* M_1 *and the numbers a and* c_1 *are computable in terms of the Taylor series of* (3) *at p.*

REFERENCES

[1] N. FENICHEL, *Oscillatory bifurcations in singular perturbation theory*, I. *Slow oscillations*, this Journal, this issue, pp. 861–867.

[2] _____, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31, (1979), pp. 53–98.

[3] B. HASSARD AND Y.-H. WAN, *Bifurcation formulas derived from center manifold theory*, J. Math. Anal. Appl., 63 (1978), pp. 297–312.

[4] N. LEBOVITZ AND R. SCHAAR, *Exchange of stabilities in autonomous systems*, Stud. Math., 14 (1975), pp. 229–259.

[5] R. SACKER, *On invariant surfaces and bifurcation of periodic solutions of ordinary differential equations*, IMM-NYU 333, New York Univ., New York, October, 1964.

[6] B. SEGRÉ, *Some Properties of Differentiable Varieties and Transformations*, Springer-Verlag, Berlin, 1957.

[7] J. SIJBRAND, *Studies in nonlinear stability and bifurcation theory*, Thesis, Utrecht, 1981.

[8] Y.-H. WAN, *On the uniqueness of invariant manifolds*, J. Differential Equations, 24 (1977), pp. 268–273.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0004 $01.25/0

SYSTEMS OF SINGULAR PERTURBATION PROBLEMS WITH A FIRST ORDER TURNING POINT*

PETER A. MARKOWICH† AND C. A. RINGHOFER‡

Abstract. This paper deals with systems of singularly perturbed ordinary differential equations with a first order turning point. Two-point boundary conditions are attached. We treat interior turning points and boundary turning points and give in each case an estimate for the norm of the inverse of the differential operator. In both cases these norms tend to infinity, as the perturbation parameter ε tends to zero. In the case of a boundary turning point, the norm of the inverse blows up algebraically, as $\varepsilon^{-1/2}$, and in the interior turning point case, the blow-up is exponential, as $\varepsilon^{-1/2}\exp(\omega/2\varepsilon)$ with $\omega>0$.

For linear and quasilinear problems we prove existence (and uniqueness) results and investigate the asymptotic behavior of the solutions as $\varepsilon\to 0$. In the boundary turning point case, we show that solutions are uniformly bounded in compact subsets of the open interval and converge there uniformly (as $\varepsilon>0$), to the solution of the reduced equation ($\varepsilon=0$). At both boundary points, layers of height $O(\varepsilon^{-1/2})$ occur generally. In the interior turning point case the solutions generally blow up exponentially (at least left or right from the turning point).

AMS-MOS subject classification (1980). Primary 34B15, 34D15, 34E05, 34C11, 34E20

Key words. nonlinear boundary value problems, singular perturbations, asymptotic expansions, boundedness of solutions, turning point theory

1. Introduction.

In this paper we investigate boundary value problems of systems of singularly perturbed differential equations with a first order turning point.

The problems we study have the form

$$\varepsilon y' = tA(t)y + \varepsilon h(y,t,\varepsilon) + f(t,\varepsilon), \qquad 0\le t\le 1, \tag{1.1}$$

$$B_0(\varepsilon)y(0,\varepsilon) + B_1(\varepsilon)y(1,\varepsilon) = \beta(\varepsilon) \tag{1.2}$$

where y, f, β are n-vectors, h: $\mathbb{R}^n\times[0,1]\times[0,\varepsilon_0]\to\mathbb{R}^n$, A, B_0, B_1 are $n\times n$ matrices and $0<\varepsilon\le\varepsilon_0\ll 1$ holds. The real parts of the eigenvalues of $A(t)$ are bounded away from zero uniformly for $t\in[0,1]$ such that $t=0$ is a first order (boundary) turning point of (1.1).

We also study interior turning point problems for linear systems

$$\varepsilon y' = tAy + f(t,\varepsilon), \qquad -1\le t\le 1, \tag{1.3}$$

$$B_{-1}(\varepsilon)y(-1,\varepsilon) + B_1(\varepsilon)y(1,\varepsilon) = \beta(\varepsilon). \tag{1.4}$$

Coupled singular perturbation problems of the form

$$\varepsilon y' = A_1(t,\varepsilon)y + A_2(t,\varepsilon)z + f_1(t,\varepsilon), \qquad 0\le t\le 1, \tag{1.5}$$

$$z' = A_3(t,\varepsilon)y + A_4(t,\varepsilon)z + g_1(t,\varepsilon), \qquad 0\le t\le 1, \tag{1.6}$$

$$F_0(\varepsilon)\begin{pmatrix} y(0,\varepsilon)\\ z(0,\varepsilon)\end{pmatrix} + F_1(\varepsilon)\begin{pmatrix} y(1,\varepsilon)\\ z(1,\varepsilon)\end{pmatrix} = \gamma(\varepsilon) \tag{1.7}$$

*Received by the editors November 30, 1981, and in revised form June 26, 1982.

†Department of Mathematics, University of Texas at Austin, Austin, Texas 78712. The research of this author was sponsored 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-7927062.

‡Institut für Angewandte Mathematik, Technische Universität Wien, Gusshausstrasse 27-29, A-1040 Vienna, Austria. The research of this author was supported by the Österreichische Forschungsförderungsverein.

are well understood if the real parts of the eigenvalues of $A_1(t,\varepsilon)$ are bounded away from zero uniformly for $t\in[0,1]$ and small ε. The perturbed n-vector y is called the fast component and the unperturbed m-vector z is called the slow component. If the coefficient matrices and the forcing terms of (1.5), (1.6) are sufficiently smooth and if (1.7) fulfills a certain regularity condition, then the unique solution y,z of (1.6), (1.7) fulfills

$$y(t,\varepsilon)=\bar{y}(t)+\sigma_-\left(\frac{t}{\varepsilon}\right)+\sigma_+\left(\frac{1-t}{\varepsilon}\right)+O(\varepsilon), \tag{1.8a}$$

$$z(t,\varepsilon)=\bar{z}(t)+O(\varepsilon) \tag{1.8b}$$

where $\|\sigma_-(\tau)\|$, $\|\sigma_+(\tau)\|\le c_0 e^{-c_1\tau}$ holds and $\bar{y},\bar{z}$ are (smooth) solutions of the reduced equations

$$0=A_1(t,0)\bar{y}(t)+A_2(t,0)\bar{z}(t)+f_1(t,0), \tag{1.9}$$

$$\bar{z}'=A_3(t,0)\bar{y}(t)+A_4(t,0)\bar{z}+g_1(t,0) \tag{1.10}$$

with m appropriate boundary conditions. The stability estimate

$$\|y(\cdot,\varepsilon)\|_{[0,1]}+\|z(\cdot,\varepsilon)\|_{[0,1]}\le\text{const}\left(\|f_1(\cdot,\varepsilon)\|_{[0,1]}+\|g_1(\cdot,\varepsilon)\|_{[0,1]}+\|\gamma(\varepsilon)\|\right) \tag{1.11}$$

holds where $\|\cdot\|_{[a,b]}$ denotes the max-norm on $[a,b]$. Proof of these statements can be found in O'Malley (1979) and Ringhofer (1981a, b).

The fast component y has (generally) a boundary layer of width $O(\varepsilon|\ln\varepsilon|)$ at $t=0$ and $t=1$, is smooth in $[c\varepsilon|\ln\varepsilon|, 1-c\varepsilon|\ln\varepsilon|]$, $c>0$, and converges to $\bar{y}(t)$ in $(0,1)$. The slow component is smooth in $[0,1]$ and converges uniformly to $\bar{z}$ in $[0,1]$.

Singular perturbation problems of the form (1.5), (1.6) defined on the infinite interval $[0,\infty]$ have been investigated by Markowich and Ringhofer (1983)(this issue, pp. 897–914). The results obtained for finite interval problems carry over if only solutions with a finite limit at $t=\infty$ are admitted (except that there is no layer at $t=\infty$). Analogously, quasilinear problems (see Ringhofer (1981a, b)) can be dealt with.

Also the numerical treatment of these non turning point problems is well understood (see Kreiss and Nichols (1975), Ringhofer (1980), (1981) and Ascher and Weiss (1981) for further references).

O'Malley (1978), (1979) investigated problems where $A_1(t,0)$ has a constant number of semisimple zero eigenvalues (called singular singularly perturbed problems).

Much work has been done on the scalar second order equation with a first order turning point:

$$\varepsilon x''+a(t)x'+b(t)x=f(t), \qquad -1\le t\le 1, \tag{1.12}$$

$$x(-1,\varepsilon)=\alpha, \quad x(1,\varepsilon)=\beta, \qquad \alpha,\beta\in\mathbb{R} \tag{1.13}$$

with $a(0)=0$, $a'(0)\neq 0$ and $a(t)\neq 0$ for $t\in[-1,0)\cup(0,1]$ (see Abrahamsson (1975), Ackerberg and O'Malley (1970), Kreiss and Parter (1971) and O'Malley (1970)). A discussion of nonlinear second order equations and a collection of references can be found in Howes (1980).

The problem (1.12), (1.13) (after transformation to a system setting $y=x'$, $z=x$) can be regarded as (the simplest) example of a coupled system with a first order turning point. The behaviour of solutions mainly depends on the sign of $a'(0)$ and on $b(0)/|a'(0)|$. Problem (1.12), (1.3) is accessible to a special function approach.

Wasow (1978) constructs asymptotic expansions for fundamental matrices of n-dimensional systems of the form

$$\varepsilon^k y' = \tilde{A}(t,\varepsilon)y, \quad |t| \le t_0, \quad 0 < \varepsilon \le \varepsilon_0, \quad k \in \mathbb{N} \tag{1.14}$$

where $\tilde{A}$ is analytic in both variables.

These expansions are valid outside boundary layer regions and no restrictions on the eigenvalues of $\tilde{A}(t,\varepsilon)$ are necessary.

Our approach to (1.1), (1.2) and (1.3), (1.4) is the following. We construct uniformly bounded fundamental matrices for

$$\varepsilon y' = tA(t)y + f(t,\varepsilon) \tag{1.15}$$

and treat the forcing term $f(t,\varepsilon)$ by variation of constants. For the quasilinear problem we employ a contraction argument. It seems likely that this approach together with Wasow's construction of fundamental matrices can be used to treat problems with more complicated eigenvalue configurations. (Our set-up assumes that every eigenvalue of $\tilde{A}(t,0)$ changes sign at $t=0$!)

In contrast to singular perturbation problems without turning points, the differential operator (1.15) (with boundary conditions) has no uniformly bounded inverse on $C([0,1])$ and $C([-1,1])$ respectively. Optimal bounds for the norm of the inverse are $c_0/\sqrt{\varepsilon}$ on $C([0,1])$ and $(c_1/\sqrt{\varepsilon})\exp(\omega/2\varepsilon)$; c_0, c_1, $\omega > 0$ on $C([-1,1])$. Interior turning point problems are exponentially unstable.

The results for the boundary turning point problem indicate that we have to distinguish two cases. First assume that $f(t,\varepsilon) = tg(t,\varepsilon)$, g sufficiently smooth. Then the solution of the reduced problem (1.1) (defined by setting $\varepsilon = 0$ in (1.1)) is $\bar{y}(t) = -A^{-1}(t)g(t,0)$ and is smooth on $[0,1]$. Assuming regularity of the boundary condition (1.2) we show that this puts us back into the non turning point case. The solution $y(t,\varepsilon)$ of (1.1), (1.2) fulfills

$$y(t,\varepsilon) = \bar{y}(t) + \sigma_-\left(\frac{t^2}{2\varepsilon}\right) + \sigma_+\left(\frac{1-t^2}{2\varepsilon}\right) + o(1), \qquad \varepsilon \to 0, \tag{1.16}$$

uniformly on $[0,1]$.

However, the smoothness of $\bar{y}$ does not guarantee uniform boundedness of solutions of the interior turning point problem (1.3), (1.4). Exponential blow-up (as $\varepsilon \to 0$) of the solution of (1.3), (1.4) (generally) occurs. Dropping the smoothness of $\bar{y}(t)$, we cannot even expect uniformly bounded solutions of (1.1), (1.2). Assuming that $f \in C^1([0,1])$ uniformly as $\varepsilon \to 0$ and that a certain growth condition on h holds, we show that

$$y(t,\varepsilon) = \bar{y}(t) + \frac{1}{\sqrt{\varepsilon}}\left(\sigma_-\left(\frac{t^2}{2\varepsilon}\right) + \sigma_+\left(\frac{1-t^2}{2\varepsilon}\right)\right) + \frac{\varepsilon}{t^3}\rho(t,\varepsilon) + o(1) \tag{1.17}$$

where $\bar{y}(t) = -A^{-1}(t)f(t,0)/t$, $\|\rho\|_{[0,1]} \le \text{const}$ and $0 < t = t(\varepsilon)$ such that $t(\varepsilon)/\sqrt{\varepsilon} \to \infty$ and $\varepsilon \to 0$. Globally on $[0,1]$ we get $\|y(t,\varepsilon)\| \le \text{const.}\, 1/\sqrt{\varepsilon}$.

Therefore the solution $y(t,\varepsilon)$ converges to $\bar{y}(t)$ on $(0,1)$ as $\varepsilon \to 0$ and the convergence is uniform on $[\bar{t}(\varepsilon), 1 - c\varepsilon|\ln\varepsilon|]$ whenever $\bar{t}(\varepsilon)/(\varepsilon)^{1/3} \to \infty$ as $\varepsilon \to 0$. $y(t,\varepsilon)$ (generally) has a boundary layer at $t=0$ and $t=1$, both being of height $O(1/\sqrt{\varepsilon})$. The layer at $t=1$ is of width $O(\varepsilon|\ln\varepsilon|)$.

The results for the boundary turning point problems (1.1), (1.2) can easily be extended to turning points of order $\alpha > 0$, i.e., $tA(t)$ in (1.1) is substituted by $t^\alpha A(t)$.

The paper is organized as follows. Section 2 deals with scalar equations; linear problems with constant coefficient matrices are dealt with in §3, (using the result of §2); in §4 we admit variable coefficients, and §5 is concerned with quasilinear problems.

2. Scalar constant coefficient problems.

A. **Coefficients with negative real parts.** At first we investigate the asymptotic behaviour of solutions of the problem

$$\varepsilon y' = -aty + f(t,\varepsilon), \quad -1 \le t \le 1 \text{ or } 0 \le t \le 1, \quad a \in \mathbb{C}, \quad \operatorname{Re} a > 0. \tag{2.1}$$

The general solution of (2.1) is given by

$$y(t,\varepsilon) = \varphi(t,\varepsilon,a)\xi + (H_\varepsilon^-(a)f)(t), \tag{2.2}$$

where $\xi \in C$, $\xi = y(0,\varepsilon)$ and

$$\varphi(t,\varepsilon,a) = \exp\left(\frac{-a}{2\varepsilon}t^2\right), \tag{2.3a}$$

$$(H_\varepsilon^-(a)f)(t) = \frac{1}{\varepsilon}\int_0^t \exp\left(\frac{a}{2\varepsilon}(s^2-t^2)\right) f(s,\varepsilon)\,ds \tag{2.3b}$$

hold. We estimate the norm of $H_\varepsilon^-(a)$ in

LEMMA 2.1. *The operator* $H_\varepsilon^-(a)$: $C([t_0,1]) \to C([t_0,1])$ *for* $t_0 = -1$ *or* $t_0 = 0$ *fulfills*

$$\| H_\varepsilon^-(a) \|_{[t_0,1]} = c(a,\varepsilon)\frac{1}{\sqrt{\varepsilon}}, \qquad \varepsilon > 0, \tag{2.4}$$

where $0 < \underline{c} \le c(a,\varepsilon) \le \bar{c}$ *for* $\varepsilon \in (0,\varepsilon_0)$ *and* $0 < a_1 \le \operatorname{Re} a \le a_2$. $\underline{c}, \bar{c}$ *only depend on* ε_0, a_1, a_2, t_0.

Proof. Choosing $f(t,\varepsilon) = \exp(-i \operatorname{Im} at^2/2\varepsilon)$ ($\| f \|_{[t_0,1]} = 1$) and evaluating at $t = \sqrt{\varepsilon}$ gives

$$(H_\varepsilon^-(a)f)(\sqrt{\varepsilon}) = e^{-a/2}\frac{1}{\varepsilon}\int_0^{\sqrt{\varepsilon}} \exp\left(\frac{\operatorname{Re} a}{2\varepsilon}s^2\right) ds$$

since

$$\int_0^{\sqrt{\varepsilon}} \exp\left(\frac{\operatorname{Re} a}{2\varepsilon}s^2\right) ds \ge \sqrt{\varepsilon}\,.$$

We obtain

$$\| H_\varepsilon^-(a) \|_{[t_0,1]} \ge e^{-(\operatorname{Re} a/2)}\frac{1}{\sqrt{\varepsilon}}\,.$$

Moreover, for general $f \in C([t_0,1])$

$$|(H_\varepsilon^-(a)f)(t)| \le \frac{1}{\varepsilon}\int_0^{|t|} \exp\left(\frac{\operatorname{Re} a}{2\varepsilon}(s^2-t^2)\right) ds \, \| f \|_{[0,t]}$$

holds. The substitutions $x = \sqrt{\operatorname{Re} a/2\varepsilon}\, s$ and $u = \sqrt{\operatorname{Re} a/2\varepsilon}\, |t|$ give

$$\frac{1}{\varepsilon}\int_0^{|t|} \exp\left(\frac{\operatorname{Re} a}{2\varepsilon}(s^2-t^2)\right) ds = \sqrt{\frac{2}{\varepsilon \operatorname{Re} a}}\, e^{-u^2}\int_0^u e^{x^2}\,dx.$$

A calculation shows that

$$e^{-u^2}\int_0^u e^{x^2}dx=\frac{1}{2u}+O\left(\frac{1}{u^3}\right)\quad \text{as } u\to\infty \tag{2.5}$$

holds, and Lemma 2.1 follows.

We investigate the existence of uniformly bounded solutions (as $\varepsilon\to 0$) of (2.1):

LEMMA 2.2. *If* $f(t,\varepsilon)=tg(t,\varepsilon)+\sqrt{\varepsilon}\,h(t,\varepsilon)$ *with* $g\in C([t_0,0)\cup(0,1])\cap L^\infty([t_0,1])$, $h\in C([t_0,1])$, $t_0=-1$ *or* $t_0=0$ *for* $0<\varepsilon\le\varepsilon_0$ *then*

$$\|H_\varepsilon^-(a)f\|_{[t_0,1]}\le \text{const}\left(\|g\|_{[t_0,1]}+\|h\|_{[t_0,1]}\right) \tag{2.6}$$

holds, where const *is independent of* $\varepsilon\in(0,\varepsilon_0]$ *and* $\operatorname{Re}a\in[a_1,a_2]$, $a_1>0$, $\operatorname{Im}a\in\mathbb{R}$. *If* $f(t,\varepsilon)\equiv f(t)\in\mathbb{R}$ *for* $t\in[t_0,1]$, $f\in C([t_0,1])$ *and if there is an* $\alpha>0$ *such that the sign of* $f(t)$ *is constant on* $(0,\alpha]$ *(and on* $[-\alpha,0)$ *if* $t_0=-1$*) then* $(H_\varepsilon^-(a)f)(t)$ *is uniformly bounded on* $[t_0,1]$ *as* $\varepsilon\to 0$ *if and only if*

$$f(t)=tg(t),\qquad g\in C([t_0,0)\cup(0,1])\cap L^\infty([t_0,1]). \tag{2.7}$$

Proof. The first statement follows from

$$|(H_\varepsilon^-(a)f)(t)|\le c\left(\frac{1}{\varepsilon}\exp\left(-\frac{\operatorname{Re}a}{2\varepsilon}t^2\right)\int_0^{|t|}\exp\left(\frac{\operatorname{Re}a}{2\varepsilon}s^2\right)s\,ds\,\|g\|_{[t_0,t]}+\|h\|_{[t_0,1]}\right).$$

Here (2.4) was used. Obviously

$$\int_0^{|t|}\exp\left(\frac{\operatorname{Re}a}{2\varepsilon}s^2\right)s\,ds=\frac{\varepsilon}{\operatorname{Re}a}\left(\exp\left(\frac{\operatorname{Re}a}{2\varepsilon}t^2\right)-1\right)$$

holds and (2.6) follows.

Now we have to show the necessity of (2.7).

We choose δ such that $0<\delta<\sqrt{\pi/|\operatorname{Im}a|}$ and ε so small that $\sqrt{\varepsilon}\,\delta<\alpha$. We get

$$\begin{aligned}\left|(H_\varepsilon^-(a)f)(\delta\sqrt{\varepsilon})\right|&\ge\frac{1}{\varepsilon}\exp\left(-\frac{\operatorname{Re}a}{2}\delta^2\right)\left|\int_{(\delta/2)\sqrt{\varepsilon}}^{\delta\sqrt{\varepsilon}}\exp\left(\frac{a}{2\varepsilon}s^2\right)f(s)\,ds\right|\\&\ge\left(\exp\left(-\frac{\operatorname{Re}a}{2}\delta^2\right)\left|\frac{f(\xi)}{\xi}\right|\right)\\&\quad\cdot\left(\frac{1}{\varepsilon}\int_{(\delta/2)\sqrt{\varepsilon}}^{\delta\sqrt{\varepsilon}}\exp\left(\frac{\operatorname{Re}a}{2\varepsilon}s^2\right)\cos\left(\frac{\operatorname{Im}a}{2\varepsilon}s^2\right)s\,ds\right)\end{aligned}$$

for some $\xi\in[\delta\sqrt{\varepsilon}/2,\ \delta\sqrt{\varepsilon}]$. The second factor is bounded from above and below as $\varepsilon\to 0$ and the first is bounded only if (2.7) holds on $[0,1]$. The same consideration holds on $[-1,0]$ (if $t_0=-1$).

If $f(t,\varepsilon)\equiv tg(t)$ with g as in Lemma 2.2, then the solution $\bar{y}$ of the reduced equation (2.1) is

$$\bar{y}(t)=\frac{1}{a}g(t)\in L^\infty([t_0,1]) \tag{2.8}$$

and if $g\in C^1([t_0,1])$ ($t_0=-1$ or $t_0=0$) we obtain (using a perturbation argument)

$$(H_\varepsilon^-(a)f)(t)=\bar{y}(t)-\varphi(t,\varepsilon,a)\bar{y}(0)+O\left(\sqrt{\varepsilon}\,\|g\|_{[t_0,1]}\right) \tag{2.9}$$

uniformly on $[t_0,1]$ and the turning point $t>0$ produces a boundary layer of width $O(\sqrt{\varepsilon})$ unless $y(0,\varepsilon)\equiv\bar{y}(0)$.

Now we turn to the case where the solution of the reduced equation $\bar{y}$ is not in L^∞ and show

LEMMA 2.3. *For $f\in C([t_0,1])$, $t_0=0$ or $t_0=-1$, the estimate*

$$|(H_\varepsilon^-(a)f)(t)|\le \text{const}\min\left(\frac{|t|}{\varepsilon},\frac{1}{|t|}\right)\|f\|_{[t_0,1]},\qquad t_0\le t\le 1, \tag{2.10}$$

holds uniformly for $\operatorname{Re}a\in[a_1,a_2]$, $a_1>0$, $\operatorname{Im}a\in\mathbb{R}$.

Proof. Let $|t|\le\sqrt{\varepsilon}$. Then

$$|(H_\varepsilon^-(a)f)(t)|\le\text{const}\frac{1}{\varepsilon}\int_0^{|t|}ds\,\|f\|_{[t_0,1]}=\text{const}\frac{|t|}{\varepsilon}\|f\|_{[t_0,1]}.$$

From the proof of Lemma 2.1 we get

$$|(H_\varepsilon^-(a)f)(t)|\le\text{const}\frac{1}{\sqrt{\varepsilon}}e^{-u^2}\int_0^u e^{x^2}dx\,\|f\|_{[t_0,1]}$$

with $u=\sqrt{\operatorname{Re}a/2\varepsilon}\,|t|$. Using (2.5) we obtain for $|t|\ge\sqrt{\varepsilon}$

$$|(H_\varepsilon^-(a)f)(t)|\le\text{const}\frac{1}{|t|}\|f\|_{[t_0,1]},$$

and (2.10) follows.

Lemma 2.3 implies that (in general) there is an interior layer at $t=0$ of thickness $O(\sqrt{\varepsilon})$ and height $O(1/\sqrt{\varepsilon})$.

For $f(t,\varepsilon)\equiv f(t)\in\mathbb{R}$

$$|(H_\varepsilon^-(a)f)(t)\ge\text{const}\frac{|t|}{\varepsilon}\min_{t\in[-\alpha,\alpha]}|f(t)|,\qquad |t|\le\sqrt{\varepsilon}, \tag{2.11}$$

holds such that $(H_\varepsilon^-(a)f)(t)$ is bounded from below and above by linear functions of slope $O(1/\varepsilon)$ (inside the interior layer).

The following lemma describes $(H_\varepsilon^-(a)f)(t)$ outside the interior layer:

LEMMA 2.4. *Let $0\ne t=t(\varepsilon)\in[t_0,1]$ ($t_0=-1$ or $t_0=0$) be such that $t(\varepsilon)/\sqrt{\varepsilon}\to\infty$ or $t(\varepsilon)/\sqrt{\varepsilon}\to-\infty$. Then if $f(t,\varepsilon)=f(t)$, $f\in C^1([t_0,1])$, the expansion*

$$(H_\varepsilon^-(a)f)(t(\varepsilon))=\bar{y}(t(\varepsilon))+\frac{\varepsilon}{t(\varepsilon)^3}x(t(\varepsilon),\varepsilon) \tag{2.12}$$

holds where $\bar{y}(t)=f(t)/at$, $t\ne0$, is the solution of the reduced equation (2.1) *and*

$$\|x\|_{[t_0,1]}\le\text{const}\left(\|f\|_{[t_0,1]}+\|f'\|_{[t_0,1]}\right) \tag{2.13}$$

uniformly for $\varepsilon\in(0,\varepsilon_0]$, $\operatorname{Re}a\in[a_1,a_2]$, $a_1>0$ and $\operatorname{Im}a\in\mathbb{R}$.

Proof. The substitutions $s=\sqrt{2\varepsilon/a}\,x$ and $T=at/2\sqrt{\varepsilon}$ give

$$(H_\varepsilon^-(a)f)(t)=\frac{1}{at}\left(\frac{2T\int_0^T e^{x^2}f(tx/T)\,dx}{e^{T^2}}\right).$$

The assumption $t(\varepsilon)/\sqrt{\varepsilon}\to\pm\infty$ implies that $T\to\infty$ (in the complex plane) and the application of Hospital's rule gives

$$(H_\varepsilon^-(a)f)(t)=\frac{f(t)}{at}+\frac{1}{tT^2}\tilde{x}(T,t)$$

where $\tilde{x}$ fulfills (2.13). Resubstitution gives (2.12).

An easy calculation shows that (2.12) also holds for $t(\varepsilon)=c\sqrt{\varepsilon}$, $c\ne0$.

B. Coefficients with positive real parts. The general solution of the problem

$$\varepsilon y' = bty + f(t,\varepsilon), \qquad -1 \le t \le 1 \text{ or } 0 \le t \le 1, \quad b \in \mathbb{C}, \quad \operatorname{Re} b > 0, \tag{2.14}$$

is

$$y(t,\varepsilon) = \psi(t,\varepsilon,b)\eta + (H_\varepsilon^+(b)f)(t) \tag{2.15}$$

where

$$\psi(t,\varepsilon,b) = \exp\left(\frac{b}{2\varepsilon}(t^2-1)\right), \tag{2.16a}$$

$$(H_\varepsilon^+(b)f)(t) = \frac{1}{\varepsilon}\int_1^t \exp\left(\frac{b}{2\varepsilon}(t^2-s^2)\right) f(s,\varepsilon)\,ds \tag{2.16b}$$

holds with $\eta = y(1,\varepsilon)$.

The following lemma, which gives estimates of the norm of $H_\varepsilon^+(b)$, should be compared with Lemma 2.1.

LEMMA 2.5. *The operator* $H_\varepsilon^+(b)$: $C([-1,1]) \to C([-1,1])$ *fulfills*

$$\|H_\varepsilon^+(b)\|_{[-1,1]} = \bar{c}(b,\varepsilon)\frac{1}{\sqrt{\varepsilon}}\exp\left(\frac{\operatorname{Re} b}{2\varepsilon}\right) \tag{2.17}$$

and, when regarded as operator from $C([0,1])$ *to* $C([0,1])$, *it fulfills*

$$\|H_\varepsilon^+(b)\|_{[0,1]} = \tilde{c}(b,\varepsilon)\frac{1}{\sqrt{\varepsilon}} \tag{2.18}$$

where $\bar{c}$ *and* $\tilde{c}$ *have the properties of the function* c *in Lemma* 2.1.

Proof. We take the function $f(t,\varepsilon) = \exp(i \operatorname{Im} bt^2/2\varepsilon)$, evaluate at $t=0$ and $t=-1$ and obtain

$$(H_\varepsilon^+(b)f)(0) = -e^{b/\varepsilon}\frac{1}{\varepsilon}\int_0^1 \exp\left(-\frac{\operatorname{Re} b}{2\varepsilon}s^2\right) ds,$$

$$(H_\varepsilon^+(b)f)(-1) = -e^{(b/2\varepsilon)}\frac{1}{\varepsilon}\int_{-1}^1 \exp\left(-\frac{\operatorname{Re} b}{2\varepsilon}s^2\right) ds.$$

A simple calculation gives

$$|(H_\varepsilon^+(b)f)(0)| \ge \text{const}\frac{1}{\sqrt{\varepsilon}},$$

$$|(H_\varepsilon^+(b)f)(-1)| \ge \text{const}\frac{1}{\sqrt{\varepsilon}}e^{(\operatorname{Re} b/2\varepsilon)}$$

where the constant is independent of ε, $\operatorname{Im} b \in \mathbb{R}$ and $\operatorname{Re} b \in [b_1, b_2]$, $b_1 > 0$. (2.17), (2.18) follow by proceeding similarly to the proof of Lemma 2.1.

Let $C_{\text{odd}}([-1,1])$ be the space of odd $C([-1,1])$-functions and $C_{\text{even}}([-1,1])$ be the space of even $C([-1,1])$-functions. Then we show

LEMMA 2.6. *The operator* $H_\varepsilon^+(b)$: $C_{\text{odd}}([-1,1]) \to C_{\text{even}}([-1,1])$ *fulfills*

$$\|H_\varepsilon^+(b)\|_{C_{\text{odd}}([-1,1])} = \tilde{c}(b,\varepsilon)\frac{1}{\sqrt{\varepsilon}} \tag{2.19}$$

where $\tilde{c}$ *is defined in Lemma* 2.5. *If* $f(t,\varepsilon)=tg(t,\varepsilon)+\sqrt{\varepsilon}\,h(t,\varepsilon)$ *with* $g\in C((0,1])\cap L^\infty([0,1])$, $h\in C([0,1])$ *then*

$$\|(H_\varepsilon^+(b)f)\|_{[0,1]}\leq \text{const}(\|g\|_{[0,1]}+\|h\|_{[0,1]}) \tag{2.20}$$

where const *is independent of* $\varepsilon\in(0,\varepsilon_0]$, *and* $\operatorname{Im} b\in\mathbb{R}$, $\operatorname{Re} b\in[b_1,b_2]$, $b_1>0$. *If* $f(t,\varepsilon)\equiv f(t)\in C([-1,1])$ *is real and analytic in* $[-1,1]$ *then* $(H_\varepsilon^+(b)f)(t)$ *is uniformly bounded on* $[-1,1]$ *if and only if*

$$f(t)=tg(t),\qquad g\in C_{\text{even}}([-1,1]). \tag{2.21}$$

Proof. (2.19) holds because the integrand in (2.16b) is odd if $f(s,\varepsilon)$ is odd and (2.20) is proven analogously to (2.6).

We only have to prove the necessity of (2.21).

We choose $\alpha>0$ sufficiently small and write

(2.22)

$$(H_\varepsilon^+(b)f)(t)=-\frac{1}{\varepsilon}\int_t^\alpha \exp\left(\frac{b}{2\varepsilon}(t^2-s^2)\right)f(s)\,ds-\frac{1}{\varepsilon}\int_\alpha^1 \exp\left(\frac{b}{2\varepsilon}(t^2-s^2)\right)f(s)\,ds$$

and set $t=0$. The second term on the right-hand side of (2.22) can be estimated by $\text{const}(1/\varepsilon)\exp(-\operatorname{Re} b\alpha^2/2\varepsilon)\|f\|_{[0,1]}$. For the first term we get

$$\left|\frac{1}{\varepsilon}\int_0^\alpha \exp\left(-\frac{bs^2}{2\varepsilon}\right)f(s)\,ds\right|\geq\frac{1}{\varepsilon}\left|\int_0^\alpha \exp\left(-\frac{\operatorname{Re} bs^2}{2\varepsilon}\right)\cos\left(\frac{\operatorname{Im} b}{2\varepsilon}s^2\right)(f(0)+sf'(\gamma))\,ds\right|$$

where $\gamma\in[0,\alpha]$. Obviously

$$\left|\frac{1}{\varepsilon}\int_0^\alpha \exp\left(-\frac{\operatorname{Re} b}{2\varepsilon}s^2\right)\cos\left(\frac{\operatorname{Im} b}{2\varepsilon}s^2\right)sf'(\gamma)\,ds\right|\leq\text{const}\|f\|_{[0,\alpha]}$$

holds. In the remaining term we substitute $x=\sqrt{\operatorname{Re} b/2\varepsilon}\,s$, $T=\sqrt{\operatorname{Re} b/2\varepsilon}\,\alpha$ and since

$$\lim_{T\to\infty}\left|\int_0^T e^{-x^2}\cos\left(\frac{\operatorname{Im} b}{\operatorname{Re} b}x^2\right)ds\right|=\infty,$$

$f(0)=0$ has to hold in order to make $H_\varepsilon^+(b)f$ bounded on $[0,1]$ as $\varepsilon\to0$. From the analyticity of f in $t=0$ we conclude that $f(t)=tg(t)$, $g\in C([-1,1])$. In order to show that $g\in C_{\text{even}}([-1,1])$, we write:

(2.23)

$$(H_\varepsilon^+(b)f)(t)=(H_\varepsilon^+(b)f)(-t)-\frac{1}{\varepsilon}\exp\left(\frac{b}{2\varepsilon}t^2\right)\int_0^t\exp\left(-\frac{b}{2\varepsilon}s^2\right)s(g(s)-g(-s))\,ds.$$

For $t<0$ the first term on the right-hand side of (2.23) uniformly bounded if and only if $f(t)=tg(t)$, $g\in C([0,1])$. Moreover

$$g(t)-g(-t)=\sum_{i=0}^\infty g_i t^{2i+1},\qquad t\in[-1,1],$$

holds. Partial integration shows that

$$\frac{1}{\varepsilon}\exp\left(\frac{b}{2\varepsilon}t^2\right)\int_0^t\exp\left(-\frac{b}{2\varepsilon}s^2\right)ss^{2i+1}\,ds$$

is for no $t<0$ (fixed) uniformly bounded (as $\varepsilon\to0$). Therefore $g_i=0$ for all i follows and g has to be even.

If $f(t,\varepsilon)=tg(t,\varepsilon)$, $g\in C^1([-1,1])$, then the following estimate is a simple consequence of (2.23):

(2.24)

$$\|H_\varepsilon^+(b)f\|_{[-1,0]}\leq \text{const}\left(\|g\|_{[-1,1]}+\sqrt{\varepsilon}\exp\left(\frac{\operatorname{Re} b}{2\varepsilon}\right)\max_{s\in[0,1]}|g'(s,\varepsilon)-g'(-s,\varepsilon)|\right).$$

The problem

$$\varepsilon y'=ty+t^2,\qquad y(1)=0, \tag{2.25}$$

has the (unbounded) solution

$$y(t,\varepsilon)=\exp\left(\frac{t^2-1}{2\varepsilon}\right)-t-\int_t^1\exp\left(\frac{t^2-s^2}{2\varepsilon}\right)ds. \tag{2.26}$$

(2.26) shows that the estimate (2.24) is sharp for this problem. However, the perturbed problem

$$\varepsilon\tilde{y}'=t\tilde{y}+(t^2-\varepsilon),\qquad \tilde{y}(1)=0, \tag{2.27}$$

has the uniformly bounded solution

$$\tilde{y}(t,\varepsilon)=\exp\left(\frac{t^2-1}{2\varepsilon}\right)-t. \tag{2.28}$$

$O(\varepsilon)$-perturbations of the inhomogeneity can produce exponentially large perturbations of the solution. Lemma 2.6 does not apply to (2.27) since the inhomogeneity depends on ε (and particularly because it changes sign on an interval of length $O(\sqrt{\varepsilon})$). As in Lemma 2.4, we get

LEMMA 2.7. *For $f\in C([-1,1])$ the estimate*

$$|(H_\varepsilon^+(b)f)(t)|\leq c\begin{cases}\min\left(\dfrac{1}{\sqrt{\varepsilon}},\dfrac{1}{t}\right)\|f\|_{[t,1]}, & t\in[0,1],\\[2ex] \dfrac{1}{\sqrt{\varepsilon}}\exp\left(\dfrac{\operatorname{Re} b}{2\varepsilon}t^2\right)\|f\|_{[t,1]}, & t\in[-1,0],\end{cases} \tag{2.29}$$

holds uniformly for $\varepsilon\in(0,\varepsilon_0]$, $\operatorname{Im} b\in\mathbb{R}$ and $\operatorname{Re} b\in[b_1,b_2]$, $b_1>0$.

We now investigate the relationship between the solution of (2.14) and the solution of the reduced equation $\bar{y}(t)=-f(t,0)/bt$.

First, assume that $f(t,\varepsilon)\equiv tg(t)$, $g\in C^1([-1,1])\cap C_{\text{even}}([-1,1])$. Then a perturbation analysis using (2.19) shows that

$$(H_\varepsilon^+(b)f)(t)=\psi(t,\varepsilon,b)\frac{1}{b}g(1)+\bar{y}(t)+O\left(\sqrt{\varepsilon}\,\|g'\|_{[0,1]}\right) \tag{2.30}$$

holds uniformly on $[-1,1]$. If g is not even (2.30) holds on $[0,1]$; however, exponential growth must be expected on $[-1,0)$.

If $\bar{y}\notin L^\infty([0,1])$ we get

LEMMA 2.8. *Let $0<t=t(\varepsilon)$ be such that $t(\varepsilon)/\sqrt{\varepsilon}\geq c>0$ as $\varepsilon\to 0$. Then if $f(t,\varepsilon)\equiv f(t)\in C^1([0,1])$ the expansion*

$$(H_\varepsilon^+(b)f)(t(\varepsilon))=\psi(t(\varepsilon),\varepsilon,b)\frac{1}{b}f(1)+\bar{y}(t(\varepsilon))+\frac{\varepsilon}{t(\varepsilon)^3}\delta(t(\varepsilon),\varepsilon) \tag{2.31}$$

holds where $\bar{y}(t) = -f(t)/bt$, $t \neq 0$, *and* $\delta(t,\varepsilon)$ *fulfills the estimate* (2.13) *with* $t_0 = 0$. *If* $f(t)$ *is odd* (2.31) *also holds for* $t(\varepsilon)/\sqrt{\varepsilon} \leq -c$.

Proof. Setting $y(t,\varepsilon) = \psi(t,\varepsilon,b)f(1)/b + \bar{y}(t) + v(t,\varepsilon)$, $y(-1,\varepsilon) = 0$, we obtain the following differential equation for v:

$$\varepsilon v' = tbv - \varepsilon \bar{y}'(t), \qquad v(1,\varepsilon) \equiv 0,$$

such that $v(t,\varepsilon) = -\varepsilon(H_\varepsilon^+(b)\bar{y}')(t)$, $t \neq 0$, holds. We calculate $\bar{y}'(t) = (f(t) - tf'(t))/t^2 b$, and since $t(\varepsilon)/\sqrt{\varepsilon} \geq c$ we get from (2.29)

$$|v(t(\varepsilon),\varepsilon)| \leq c_1 \frac{\varepsilon}{t(\varepsilon)} \|\bar{y}'\|_{[t(\varepsilon),1]} \leq c_1 \frac{\varepsilon}{t(\varepsilon)^3} \left(\|f\|_{[0,1]} + \|f'\|_{[0,1]}\right).$$

This lemma tells us that $(H_\varepsilon^+(b)f)(t)$ converges to $\bar{y}(t)$ as $\varepsilon \to 0$ pointwise on $(0,1)$ and uniformly outside the interior layer at $t > 0$ (of width $O(\sqrt{\varepsilon})$) and the boundary layer at $t = 1$ (of width $O(\varepsilon|\ln \varepsilon|)$).

C. Behaviour of derivatives of solutions. Now we give estimates of the derivatives of the solutions of (2.1) and (2.14) which are important for the analysis of finite difference methods.

LEMMA 2.9. *Let* $f \in C^k([-1,1])$. *Then*

$$(2.32) \quad \left|(H_\varepsilon^-(a)f)^{(k)}(t)\right| \leq \text{const}\left(\varepsilon^{-k}\exp\left(-\frac{\text{Re}\, a}{2\varepsilon}t^2\right) + \min(\varepsilon^{-(k+1)/2}, t^{-k-1})\right) \sum_{i=0}^{k} \|f^{(i)}(\cdot,\varepsilon)\|_{[-1,1]}$$

for $t \in [-1,1]$ *and*

$$(2.33) \quad \left|(H_\varepsilon^+(b)f)^{(k)}(t)\right| \leq \text{const}\left(\varepsilon^{-k}\exp\left(\frac{\text{Re}\, b}{2\varepsilon}(t^2-1)\right) + \min(\varepsilon^{-(k+1)/2}, t^{-k-1})\right) \sum_{i=0}^{k} \|f^{(i)}(\cdot,\varepsilon)\|_{[0,1]}$$

holds for $t \in [0,1]$. $(H_\varepsilon^+(b)f)^{(k)}(t)$ *generally grows exponentially on* $[-1,0)$ *as* $\varepsilon \to 0$.

If $f(t,\varepsilon) = tg(t,\varepsilon)$, $g \in C^k([-1,1])$, $k > 0$, *then for* $t \in [-1,1]$

$$(2.34) \quad \left|(H_\varepsilon^-(a)f)^{(k)}(t)\right| \leq \text{const}\left(\varepsilon^{-k}\exp\left(-\frac{\text{Re}\, a}{2\varepsilon}t^2\right) + \min(\varepsilon^{-(k-1)/2}, t^{-k+1})\right) \sum_{i=0}^{k} \|g^{(i)}(\cdot,\varepsilon)\|_{[-1,1]},$$

and for $t \in [0,1]$,

$$(2.35) \quad \left|(H_\varepsilon^+(b)f)^{(k)}(t)\right| \leq \text{const}\left(\varepsilon^{-k}\exp\left(\frac{\text{Re}\, b}{2\varepsilon}(t^2-1)\right) + \min(\varepsilon^{-(k-1)/2}, t^{-k+1})\right) \sum_{i=0}^{k} \|g^{(i)}(\cdot,\varepsilon)\|_{[0,1]}$$

holds.

3. Constant coefficient systems.

A. Diagonalization and properties of solution operators. We consider the system of differential equations

$$\varepsilon y' = tAy + f(t,\varepsilon) \tag{3.1}$$

where A is a constant $n\times n$-matrix.

For $t\in[-1,1]$ we impose the boundary condition

$$B_{-1}(\varepsilon)y(-1,\varepsilon)+B_1(\varepsilon)y(1,\varepsilon)=\beta(\varepsilon), \tag{3.2a}$$

or for $t\in[0,1]$ we impose

$$B_0(\varepsilon)y(0,\varepsilon)+B_1(\varepsilon)y(1,\varepsilon)=\beta(\varepsilon), \tag{3.2b}$$

respectively, where $B_{-1}(\varepsilon)$, $B_0(\varepsilon)$, $B_1(\varepsilon)$ are $n\times n$-matrices, and $\beta(\varepsilon)\in\mathbb{R}^n$, $f\in C([-1,1])$ or $f\in C([0,1])$ uniformly for $\varepsilon\in[0,\varepsilon_0]$ is assumed.

In the case (3.2a) the (first order) turning point $t=0$ is in the interior of the interval considered, and in the case (3.2b) the turning point is on the boundary. We assume that A has no eigenvalue on the imaginary axis such that

$$A=EJE^{-1}, \qquad J=\begin{bmatrix} \underbrace{J_+}_{r_+} & 0 \\ 0 & \underbrace{J_-}_{r_-} \end{bmatrix}, \tag{3.3}$$

where the eigenvalues of $J_+(J_-)$ have positive (negative) real parts. We substitute

$$u=E^{-1}y \tag{3.4}$$

and obtain from (3.1)

$$\varepsilon u' = tJu + E^{-1}f(t,\varepsilon). \tag{3.5}$$

Therefore $u, E^{-1}f$ split up into u_+, $(E^{-1}f)_+\in\mathbb{C}^{r_+}$ and u_-, $(E^{-1}f)_-\in\mathbb{C}^{r_-}$. The two systems obtained from (3.5) are

$$\varepsilon u_+' = tJ_+u_+ + \big(E^{-1}f(t,\varepsilon)\big)_+, \tag{3.6a}$$

$$\varepsilon u_-' = tJ_-u_- + \big(E^{-1}f(t,\varepsilon)\big)_-. \tag{3.6b}$$

We get

$$u_+(t,\varepsilon)=\exp\left(\frac{t^2-1}{2\varepsilon}J_+\right)\xi_+ + \big(H_\varepsilon^+\big(E^{-1}f(\cdot,\varepsilon)\big)_+\big)(t), \qquad \xi_+\in\mathbb{C}^{r_+}, \tag{3.7a}$$

$$u_-(t,\varepsilon)=\exp\left(\frac{t^2}{2\varepsilon}J_-\right)\xi_- + \big(H_\varepsilon^-\big(E^{-1}f(\cdot,\varepsilon)\big)_-\big)(t), \qquad \xi_-\in\mathbb{C}^{r_-}, \tag{3.7b}$$

where

$$\big(H_\varepsilon^+h_+(\cdot,\varepsilon)\big)(t)=\frac{1}{\varepsilon}\int_1^t\exp\left(\frac{t^2-s^2}{2\varepsilon}J_+\right)h_+(s,\varepsilon)\,ds, \qquad h_+\in C([-1,1]), \tag{3.8a}$$

$$\big(H_\varepsilon^-h_-(\cdot,\varepsilon)\big)(t)=\frac{1}{\varepsilon}\int_0^t\exp\left(\frac{t^2-s^2}{2\varepsilon}J_-\right)h_-(s,\varepsilon)\,ds, \qquad h_-\in C([-1,1]), \tag{3.8b}$$

holds.

Our goal is to transform (3.6) into a system with a diagonal coefficient matrix. Proceeding similarly to Ascher and Weiss (1981), we prove

LEMMA 3.1. *Let $z_+(t,\varepsilon)$, $z_-(t,\varepsilon)$ fulfill*

$$\varepsilon z_+' = tJ_+ z_+ + h_+(t,\varepsilon), \qquad t\in[-1,1], \tag{3.9a}$$

$$z_+(1)=\xi_+\in\mathbb{C}^{r_+} \tag{3.9b}$$

and

$$\varepsilon z_-' = tJ_- z_- + h_-(t,\varepsilon), \qquad t\in[-1,1], \tag{3.10a}$$

$$z_-(0)=\xi_-\in\mathbb{C}^{r_-}. \tag{3.10b}$$

Then

$$z_+(t,\varepsilon)=\frac{1}{2\pi i}\int_{\Gamma_+} w_+(t,\varepsilon,\mu)\,d\mu, \tag{3.11}$$

$$z_-(t,\varepsilon)=\frac{1}{2\pi i}\int_{\Gamma_-} w_-(t,\varepsilon,\lambda)\,d\lambda \tag{3.12}$$

hold where the curves Γ_+ and Γ_- lie in the right and left half planes respectively and contain all eigenvalues of J_+ and J_- respectively. w_+, w_- fulfill

$$\varepsilon w_+' = t\mu w_+ + (\mu - J_+)^{-1} h_+(t,\varepsilon), \qquad t\in[-1,1], \quad \mu\in\Gamma_+, \tag{3.13a}$$

$$w_+(1,\varepsilon,\mu)=(\mu-J_+)^{-1}\xi_+ \tag{3.13b}$$

and

$$\varepsilon w_-' = t\lambda w_- + (\lambda - J_-)^{-1} h_-(t,\varepsilon), \qquad t\in[-1,1], \quad \lambda\in\Gamma_-, \tag{3.14a}$$

$$w_-(0,\varepsilon,\lambda)=(\lambda-J_-)^{-1}\xi_-. \tag{3.14b}$$

Proof. The solution of (3.9) is given by

$$z_+(t,\varepsilon)=\exp\left(\frac{t^2-1}{2\varepsilon}J_+\right)\xi_+ + (H_\varepsilon^+ h_+(\cdot,\varepsilon))(t). \tag{3.15}$$

From Dunford and Schwartz (1957, Chap. 7) we obtain the formula

$$\exp(\tau A)=\frac{1}{2\pi i}\int_{\Gamma_A} e^{\tau\gamma}(\gamma - A)^{-1}\,d\gamma, \qquad \tau\in\mathbb{R},$$

where A is a $k\times k$-matrix and Γ_A contains all eigenvalues of A. Using this formula we get

$$z_+(t,\varepsilon)=\frac{1}{2\pi i}\int_{\Gamma_+}\left(\exp\left(\frac{t^2-1}{2\varepsilon}\mu\right)(\mu-J_+)^{-1}\xi_+ + \frac{1}{\varepsilon}\int_1^t \exp\left(\frac{t^2-s^2}{2\varepsilon}\mu\right)(\mu-J_+)^{-1}h_+(s,\varepsilon)\,ds\right)d\mu. \tag{3.16}$$

Calling the integrand $w_+(t,\varepsilon,\mu)$ we see that w_+ fulfills (3.13a, b). The proof for z_- is analogous.

Obviously $\max_{\mu\in\Gamma_+}\|(\mu-J_+)^{-1}\|$, $\max_{\lambda\in\Gamma_-}\|(\lambda-J_-)^{-1}\|\le$ const holds. Using the estimates of §2 which are formulated uniformly for coefficients, which vary in bounded

sets such that their real parts are bounded away from zero, we get, collecting the results:

LEMMA 3.2. (i) *The operators* H_ε^+, H_ε^-: $C([-1,1]) \to C([-1,1])$ *fulfill*

$$\|H_\varepsilon^-\|_{[-1,1]} = c_1(\varepsilon)\frac{1}{\sqrt{\varepsilon}}, \tag{3.17}$$

$$\|H_\varepsilon^+\|_{[-1,1]} = c_2(\varepsilon)\frac{1}{\sqrt{\varepsilon}}\exp\left(\frac{\omega}{2\varepsilon}\right), \qquad \omega>0, \tag{3.18}$$

and H_ε^+: $C([0,1)] \to C([0,1])$ *fulfills*

$$\|H_\varepsilon^+\|_{[0,1]} = c_3(\varepsilon)\frac{1}{\sqrt{\varepsilon}} \tag{3.19}$$

where $0<\underline{c}\le c_1(\varepsilon), c_2(\varepsilon), c_3(\varepsilon)\le \bar{c}$ *holds for* $\varepsilon\in(0,\varepsilon_0]$.

(ii) *If*

$$f(t,\varepsilon) = tg(t,\varepsilon) + \sqrt{\varepsilon}\,h(t,\varepsilon) \tag{3.20}$$

with $g\in C([-1,0)\cup(0,1])\cap L^\infty([-1,1])$ *and* $h\in C([-1,1])$ *then*

$$\|H_\varepsilon^-(E^{-1}f)_-\|_{[-1,1]} \le \text{const}\left(\|g\|_{[-1,1]} + \|h\|_{[-1,1]}\right), \tag{3.21}$$

$$\|H_\varepsilon^+(E^{-1}f)_+\|_{[0,1]} \le \text{const}\left(\|g\|_{[0,1]} + \|h\|_{[0,1]}\right) \tag{3.22}$$

hold.

(iii) *For* $f\in C([-1,1])$ *the estimates*

$$\|\left(H_\varepsilon^-(E^{-1}f)_-\right)(t)\| \le \text{const}\,\min\left(\frac{|t|}{\varepsilon}, \frac{1}{|t|}\right)\|f\|_{[-1,1]}, \tag{3.23}$$

$$\|\left(H_\varepsilon^+(E^{-1}f)_+\right)(t)\| \le \text{const}\begin{cases} \min\left(\dfrac{1}{\sqrt{\varepsilon}}, \dfrac{1}{t}\right)\|f\|_{[t,1]}, & t\ge 0, \\ \dfrac{1}{\sqrt{\varepsilon}}\exp\left(\dfrac{\omega t^2}{2\varepsilon}\right)\|f\|_{[t,1]}, & t\le 0, \end{cases} \tag{3.24}$$

hold.

Proof. The only statement that remains to be shown is that the functions $c_1(\varepsilon)$, $c_2(\varepsilon)$, $c_3(\varepsilon)$ are bounded away from zero. We set

$$h_-(t,\varepsilon) = \sigma(t,\varepsilon)e$$

where e is a normed eigenvector of J_- to some eigenvalue γ and σ is a scalar function. We obtain

$$\begin{aligned}(H_\varepsilon^- h_-(\cdot,\varepsilon))(t) &= \frac{1}{2\pi i}\int_{\Gamma_-}\int_0^t \exp\left(\frac{t^2-s^2}{2\varepsilon}\lambda\right)(\lambda-\gamma)^{-1}\sigma(s,\varepsilon)\,ds\,d\lambda \\ &= \frac{1}{\varepsilon}\int_0^t \exp\left(\frac{t^2-s^2}{2\varepsilon}\gamma\right)\sigma(s,\varepsilon)\,ds.\end{aligned}$$

(3.17) follows by choosing $\sigma(t,\varepsilon)$ as in the proof of Lemma 2.1, and (3.18), (3.19) are proven analogously. The statements made on odd f (in §2) and Lemma 2.9 carry over too.

The asymptotics for $\varepsilon \to 0$ follow as in Lemmas 2.4, 2.8.

LEMMA 3.3. *Let* $0<t=t(\varepsilon)$ *be such that* $t(\varepsilon)/\sqrt{\varepsilon} \to +\infty$ *as* $\varepsilon \to 0$. *Then if* $f(t,\varepsilon)\equiv f(t) \in C^1([0,1])$ *the expansion*

$$(3.25)\quad \left(H_\varepsilon E^{-1} f\right)(t) = \begin{bmatrix} \exp\left(\frac{t^2-1}{2\varepsilon} J_+\right) J_+^{-1} & 0 \\ 0 & 0 \end{bmatrix} E^{-1} f(1) + \bar{u}(t) + \frac{\varepsilon}{t^3}\rho(t,\varepsilon)$$

holds where

$$H_\varepsilon = \begin{pmatrix} H_\varepsilon^+ \\ H_\varepsilon^- \end{pmatrix}, \qquad \bar{u}(t) = J^{-1}E^{-1}\frac{f(t)}{t}, \qquad t \neq 0$$

is the solution of the reduced equation (3.5) *and* ρ *fulfills*

$$(3.26)\qquad \|\rho\|_{[0,1]} \leq \text{const}\left(\|f\|_{[0,1]} + \|f'\|_{[0,1]}\right).$$

If $f(t,\varepsilon)\equiv tg(t)$, $g \in C^1([0,1])$, we obtain

(3.27)

$$\left(H_\varepsilon E^{-1} f\right)(t) = \bar{u}(t) + \begin{bmatrix} \exp\left(\frac{t^2-1}{2\varepsilon} J_+\right) & 0 \\ 0 & \exp\left(\frac{t^2}{2\varepsilon} J_-\right) \end{bmatrix} J^{-1}E^{-1} f(t) + O\left(\sqrt{\varepsilon}\,\|g'\|_{[0,1]}\right)$$

uniformly on $[0,1]$ ($\bar{u} \in L^\infty([0,1])$).

If $f(t)$ is odd (3.25) also holds on $[-1,0)$ and (3.27) holds uniformly on $[-1,1]$.

B. Interior turning point problems. We can solve (3.1) with the boundary condition (3.2a) by inserting (3.7) into (3.2a). We get

$$(3.28)\quad \left(B_{-1}(\varepsilon) + B_1(\varepsilon)\right) E \begin{bmatrix} I & 0 \\ 0 & \exp\left(\frac{J_-}{2\varepsilon}\right) \end{bmatrix} \begin{pmatrix} \xi_+ \\ \xi_- \end{pmatrix}$$
$$= \beta(\varepsilon) - B_{-1}(\varepsilon) E\left(H_\varepsilon E^{-1} f\right)(-1) - B_1(\varepsilon) E\left(H_\varepsilon E^{-1} f\right)(1).$$

The linear system (3.28) is uniquely solvable (for $\xi_+(\varepsilon), \xi_-(\varepsilon)$) if and only if $(B_{-1}(\varepsilon) + B_1(\varepsilon))^{-1}$ exists for $\varepsilon \in (0, \varepsilon_0]$.

However, if $r_- \neq 0$, uniformly bounded solutions (or even solutions which only exhibit layer behaviour) of (3.1), (3.2a) do not exist for all $\beta(\varepsilon) \in \mathbb{R}^n$ since (generally) $(H_\varepsilon^+ (E^{-1} f)_+)(t)$ blows up exponentially on $[-1,0)$ as $\varepsilon \to 0$ and since the basic solutions of (3.1) belonging to eigenvalues of J_- force $\xi_-(\varepsilon)$ to increase exponentially as $\varepsilon \to 0$ (unless the right-hand side of (3.28) is exponentially small). If $r_- = 0$ (all eigenvalues of A have positive real parts) and $\|(B_{-1}(\varepsilon) + B_1(\varepsilon))^{-1}\| \leq \text{const}$ as $\varepsilon \to 0$, then solutions which are uniformly bounded in $[-1, -\delta_1] \cup [\delta_1, 1]$, $\delta_1 > 0$, exist for all $\beta(\varepsilon)$ uniformly bounded in ε and all $f \in C_{\text{odd}}([-1,1])$ uniformly in ε. An interior layer at $t=0$ of width $O(1/\sqrt{\varepsilon})$ occurs. More generally, if $(H_\varepsilon f(\cdot,\varepsilon))(t)$ is uniformly bounded on $[-1, -\delta_1]$, then y is uniformly bounded on $[-1, -\delta_1] \cup [\delta_1, 1]$.

We collect the results in

THEOREM 3.1. (i) *The boundary value problem* (3.1), (3.2*a*) *has a unique solution for all* $\varepsilon \in \mathbb{R}^n$, $f \in C([-1,1])$, *if and only if* $(B_{-1}(\varepsilon) + B_1(\varepsilon))$ *is nonsingular.*

(ii) *If* $B_{-1}(\varepsilon)E_+ \equiv 0$ *for* $\varepsilon \in (0, \varepsilon_0]$ *then the solution splits up into*

$$y(t,\varepsilon) = \sigma_+\left(\frac{1-t^2}{2\varepsilon}\right) + \sigma_-\left(\frac{t^2-1}{2\varepsilon}\right) + h(t,\varepsilon)$$

where $h(t,\varepsilon)$ *fulfills the estimate* (3.24).

(iii) *If* f *fulfills* (3.39) *and if* $f \in C^1([0,1])$ *uniformly in* ε *and* $0 < t \le t(\varepsilon)$ *such that* $t(\varepsilon)/\sqrt{\varepsilon} \to \infty$ *holds as* $\varepsilon \to 0$, *we get*

$$h(t,\varepsilon) = \bar{y}(t) + \frac{\varepsilon}{t^3}\rho(t,\varepsilon) + O(\varepsilon)$$

where $\bar{y}(t) = -A^{-1}f(t,0)/t$ *solves the reduced equation* (3.1) *and* ρ *fulfills* (3.26).

(iv) *If* $B_1(\varepsilon)E_+ \equiv 0$ *for* $\varepsilon \in (0, \varepsilon_0]$ *we get results analogous to those in* (ii), (iii) *by interchanging* $[-1,0]$ *and* $[0,1]$.

C. Boundary turning point problems. Now we impose the boundary condition (3.2b). From (3.7) we get

$$\left(B_0(\varepsilon)E\begin{bmatrix}\exp(-J_+/2\varepsilon) & 0\\ 0 & I\end{bmatrix} + B_1(\varepsilon)E\begin{bmatrix}I & 0\\ 0 & \exp(J_-/2\varepsilon)\end{bmatrix}\right)\begin{pmatrix}\xi_+\\ \xi_-\end{pmatrix} = \beta(\varepsilon) - B_0(\varepsilon)E\big(H_\varepsilon E^{-1}f\big)(0) - B_1(\varepsilon)E\big(H_\varepsilon E^{-1}f\big)(1). \tag{3.29}$$

We assume that $B_0, B_1 \in C([0,\varepsilon_0])$ and partition

$$E = [\underbrace{E_+}_{r_+}, \underbrace{E_-}_{r_-}]. \tag{3.30}$$

If ε is sufficiently small and if

$$C = [B_1(0)E_\varepsilon, B_0(0)E_-] \tag{3.31}$$

is nonsingular, then the system (3.29) is uniquely solvable, since its coefficient matrix equals $C + O(\exp(-\delta/\varepsilon))$, $\delta > 0$, and ξ_+, ξ_- depend uniformly continuous (as $\varepsilon \to 0$) on the right-hand side.

First, assume that

$$f(t,\varepsilon) = tg(t,\varepsilon) + \varepsilon h(t,\varepsilon), \qquad g \in C((0,1]) \cap L^\infty([0,1]), \quad h \in C([0,1]), \tag{3.32}$$

uniformly as $\varepsilon \to 0$ and

$$\|g(t,\varepsilon) - g(t,0)\| = O(\varepsilon), \qquad t \in [0,1]. \tag{3.33}$$

Then $(H_\varepsilon E^{-1}f)(t)$ is uniformly bounded in $[0,1]$ as $\varepsilon \to 0$ and (3.31) guarantees that

$$\|\xi_+(\varepsilon)\|, \|\xi_-(\varepsilon)\| \le \text{const}\big(\|g\|_{[0,1]} + \sqrt{\varepsilon}\,\|h\|_{[0,1]} + \|\beta(\varepsilon)\|\big). \tag{3.34}$$

Therefore (3.7) and (3.21), (3.22) imply

$$\|y(\cdot,\varepsilon)\|_{[0,1]} \le \text{const}\big(\|\beta(\varepsilon)\| + \| + \|g(\cdot,\varepsilon)\|_{[0,1]} + \sqrt{\varepsilon}\,\|h(\cdot,\varepsilon)\|_{[0,1]}\big). \tag{3.35}$$

We assume that $g \in C^1([0,1])$ uniformly as $\varepsilon \to 0$, set

$$\bar{y}(t) = -A^{-1}g(t,0) \tag{3.36}$$

and obtain by proceeding as in (2.9)

$$y(t,\varepsilon) = \bar{y}(t) + \sigma_-\left(\frac{t^2}{2\varepsilon}\right) + \sigma_+\left(\frac{1-t^2}{2\varepsilon}\right) + O\big(\sqrt{\varepsilon}\,(\|g'\|_{[0,1]} + \|h(\cdot,\varepsilon)\|_{[0,1]})\big) \tag{3.37}$$

where $\sigma_-, \sigma_+ \in C^\infty$ are boundary layer terms fulfilling

$$\|\sigma_-^{(i)}(\tau)\|, \|\sigma_+^{(i)}(\tau)\| \leq c_i e^{-\lambda_+ \tau}, \qquad c_i, \lambda_+ > 0, \quad i \in \mathbb{N}_0. \tag{3.38}$$

Assumption (3.32) puts us back into the nonturning point case. The solution $y(t,\varepsilon)$ converges to the solution of the reduced problem $\bar{y}(t)$ in $(0,1)$ (uniformly outside the boundary layers at $t=0$ and $t=1$ of width $O(\sqrt{\varepsilon})$ and $O(\varepsilon|\ln\varepsilon|)$ respectively). The expansion (3.37) is analogous to the case where the right-hand side of the differential equation has only strictly stable and strictly unstable eigenvalues (Ringhofer (1981)).

Now we drop assumption (3.32) and assume only $f \in C([0,1])$ uniformly as $\varepsilon \to 0$ and

$$\|f(t,\varepsilon) - f(t,0)\| = 0(\varepsilon), \qquad t \in [0,1]. \tag{3.39}$$

Generally no better estimate than

$$\|(H_\varepsilon E^{-1} f)(t)\| \leq \text{const} \min\left(\frac{1}{\sqrt{\varepsilon}}, \frac{1}{t}\right) \|f\|_{[0,1]}, \qquad t \in [0,1], \tag{3.40}$$

(see (3.23), (3.24)) can be given.

From (3.29) we get:

$$\|\xi_+(\varepsilon)\|, \|\xi_-(\varepsilon)\| \leq \text{const}\left(\|\beta(\varepsilon)\| + \frac{1}{\sqrt{\varepsilon}} \|f\|_{[0,1]}\right), \tag{3.41}$$

and (3.7) gives the estimate

(3.42)

$$\|y(t,\varepsilon)\| \leq \text{const}\left[\|\beta(\varepsilon)\| + \left(\frac{1}{\sqrt{\varepsilon}}\left(\exp\left(\lambda_+ \frac{t^2-1}{2\varepsilon}\right) + \exp\left(-\lambda_+ \frac{t^2}{2\varepsilon}\right)\right) + \min\left(\frac{1}{\sqrt{\varepsilon}}, \frac{1}{t}\right)\right) \|f(\cdot,\varepsilon)\|_{[0,1]}\right]$$

for some $\lambda_+ > 0$. Therefore $y(t,\varepsilon)$ has a boundary layer of width $O(\sqrt{\varepsilon})$ and height $O(1/\sqrt{\varepsilon})$ at $t=0$ and a boundary layer of width $O(\varepsilon|\ln\varepsilon|)$ and height $O(1/\sqrt{\varepsilon})$ at $t=1$. On $[\delta_1, 1-\delta_1]$, $\delta_1 > 0$, $y(t,\varepsilon)$ is uniformly bounded as $\varepsilon \to 0$.

The turning point $t=0$ "pollutes" the solution at most at $t=0$ and $t=1$.

If $B_0(\varepsilon)E_+ \equiv 0$ then (3.29) implies that $\|\xi_+(\varepsilon)\| = O(1)$ as $\varepsilon \to 0$ and the factor $1/\sqrt{\varepsilon}$ of the exponentially decaying terms in (3.42) drops out and therefore the boundary layer at $t=1$ has height $O(1)$. Blow-up occurs at $t=0$.

We get:

THEOREM 3.2. (i) *Assume that $f \in C^1([0,1])$ uniformly in ε and that f fulfills* (3.39). *Then, if* (3.37) *holds, the boundary value problem* (3.1), (3.2b) *has a unique solution $y(t,\varepsilon)$ which fulfills*

$$y(t,\varepsilon) = \bar{y}(t) + \frac{1}{\sqrt{\varepsilon}}\left(\sigma_-\left(\frac{t^2}{2\varepsilon}\right) + \sigma_+\left(\frac{1-t^2}{2\varepsilon}\right)\right) + O\left(\sqrt{\varepsilon}\right) + \frac{\varepsilon}{t^3}\rho(t,\varepsilon), \qquad t \neq 0,$$

for $0 < t = t(\varepsilon)$ and $t(\varepsilon)/\sqrt{\varepsilon} \to +\infty$ as $\varepsilon \to 0$. ρ fulfills the estimate

$$\|\rho\|_{[0,1]} \leq \text{const}\left(\|f(\cdot,0)\|_{[0,1]} + \|f'(\cdot,0)\|_{[0,1]}\right). \tag{3.43}$$

(ii) *The estimate* (3.42) *holds for all* $t\in[0,1]$. *If* $f(t,\varepsilon)$ *fulfills* (3.32), *then the solution* y *is uniformly bounded on* $[0,1]$ *as* $\varepsilon\to 0$ *and* (3.37) *holds.*

(iii) *From Lemma* 2.9 *we get for* $f\in C^k([0,1])$

$$\|y^{(k)}(t,\varepsilon)\|\le\text{const}\left(\varepsilon^{-k-(1/2)}\left(\exp\left(-\lambda_+\frac{t^2}{2\varepsilon}\right)+\exp\left(\lambda_+\frac{t^2-1}{2\varepsilon}\right)\right)\left(\|\beta\|+\|f(\cdot,\varepsilon)\|_{[0,1]}\right)\right.$$
$$\left.+\min\left(\varepsilon^{-(k+1/2)},t^{-k-1}\right)\sum_{i=0}^{k}\|f^{(i)}(\cdot,\varepsilon)\|_{[0,1]}\right).$$

Lemma 3.7 shows that problems where the turning point is on the boundary are more stable than interior turning point problems (since no exponential blow-up can occur if the data are uniformly bounded as $\varepsilon\to 0$.

4. Variable coefficient problems. In this section we investigate the boundary turning point problem

$$\varepsilon y'=(tA(t)+\varepsilon B(t,\varepsilon))y+f(t,\varepsilon),\qquad 0\le t\le 1, \tag{4.1}$$

$$B_0(\varepsilon)y(0,\varepsilon)+B_1(\varepsilon)y(1,\varepsilon)=\beta(\varepsilon) \tag{4.2}$$

where A,B,B_0,B_1 are $n\times n$-matrices, f: $[0,1]\times[0,\varepsilon_0]\to\mathbb{R}^n$, $\beta\in\mathbb{R}^n$. We assume that

$$B_0,B_1,\beta\in C([0,\varepsilon_0]), \tag{4.3a}$$

$$B,f\in C([0,1])\quad\text{uniformly in }\varepsilon \tag{4.3b}$$

and that there is a continuous reduction of A to block form,

$$A(t)\equiv E(t)J(t)E^{-1}(t),\qquad E,E^{-1}\in C^1([0,1]), \tag{4.4}$$

such that

$$J(t)=\begin{bmatrix}\underbrace{J_+(t)}_{r_+} & \underbrace{0}_{r_-}\\ 0 & J_-(t)\end{bmatrix} \tag{4.5}$$

holds where the eigenvalues $\lambda_+(t)$ and $\lambda_-(t)$ of $J_+(t)$ and $J_-(t)$ respectively fulfill

$$\operatorname{Re}\lambda_+(t)\ge\delta_+,\quad \operatorname{Re}\lambda_-(t)\le-\delta_+,\quad t\in[0,1],\quad \delta_+>0. \tag{4.6}$$

The transformation matrix $E(t)$ exists locally if the eigenvalues of $A(t)$ split into two groups (4.6) and if $A\in C^1([0,1])$; however, the assumption of global splitting is much more restrictive (see O'Malley (1978)).

We substitute

$$y(t,\varepsilon)=E(t)u(t,\varepsilon) \tag{4.7}$$

and get from (4.1)

$$\varepsilon u'=\left(tJ(t)+\varepsilon\left(E^{-1}(t)B(t,\varepsilon)E(t)-E^{-1}(t)E'(t)\right)\right)u+E^{-1}(t)f(t,\varepsilon),\qquad 0\le t\le 1. \tag{4.8}$$

At first we investigate the perturbed system

$$\varepsilon v'=tJ(t)v+h(t,\varepsilon),\qquad 0\le t\le 1, \tag{4.9}$$

which splits up into

$$\varepsilon v'_+=tJ_+(t)v_+\varepsilon h_+(t,\varepsilon), \tag{4.10a}$$

$$\varepsilon v'_-=tJ_-(t)v_-+h_-(t,\varepsilon) \tag{4.10b}$$

where

$$v=\begin{pmatrix} v_+ \\ v_- \end{pmatrix} \quad \text{and} \quad h=\begin{pmatrix} h_+ \\ h_- \end{pmatrix}$$

have been set.

We prove:

LEMMA 4.1. *The fundamental matrix* $\phi_+(t,\varepsilon)$ *with* $\phi_+(1,\varepsilon)\equiv I_r$ *of the homogeneous problem* (4.10a) *fulfills*

$$\|\phi_+(t,\varepsilon)\phi_+^{-1}(s,\varepsilon)\| \le c_0 \exp\left(\lambda_+ \frac{t^2-s^2}{2\varepsilon}\right), \qquad 0\le t\le s\le 1, \tag{4.11}$$

and the fundamental matrix $\phi_-(t,\varepsilon)$ *with* $\phi_-(0,\varepsilon)\equiv I_{r_-}$ *of the homogeneous problem* (4.10b) *fulfills*

$$\|\phi_-(t,\varepsilon)\phi_-^{-1}(s,\varepsilon)\| \le c_0 \exp\left(-\lambda_+ \frac{t^2-s^2}{2\varepsilon}\right), \qquad 0\le s\le t\le 1, \tag{4.12}$$

where c_0, λ_+ *are positive constants and* $\varepsilon>0$ *is sufficiently small.*

Proof. Consider the homogeneous problem

$$\varepsilon z' = tJ_-(t)z, \qquad 0\le t\le 1.$$

We set $\tau=t^2/2$ and get

$$\varepsilon w'(\tau) = J_-\left(\sqrt{2\tau}\right)w(\tau), \qquad 0\le\tau\le\tfrac{1}{2}, \tag{4.13}$$

where $w(\tau)=z(t(\tau))$. The eigenvalues of $J_-(\sqrt{2\tau})$ are given by $\lambda_-(\sqrt{2\tau})$ (see (4.6)) and have strictly negative real parts on $0\le\tau\le\frac{1}{2}$. Therefore the singular perturbation problem (4.13) has strictly stable eigenvalues, and its fundamental matrix $\psi_-(\tau,\varepsilon)$ (with $\psi_-(0,\varepsilon)=I_{r_-}$) fulfills

$$\|\psi_-(\tau,\varepsilon)\| \le c_0 \exp\left(-\lambda_+ \frac{\tau}{\varepsilon}\right), \qquad 0\le\tau\le\tfrac{1}{2},$$

$$\|\psi_-(\tau,,\varepsilon)\psi_-^{-1}(\sigma,\varepsilon)\| \le c_0 \exp\left(-\lambda_+ \frac{\tau-\sigma}{\varepsilon}\right), \qquad 0\le\sigma\le\tau\le\tfrac{1}{2},$$

c_0, $c_1>0$ (see Turritin (1952)). Resubstituting $\tau=t^2/2$, $s=\sigma^2/2$ gives (4.12). Formula (4.11) follows analogously.

The solution of (4.9) can be written as:

$$v(t,\varepsilon)=\phi(t,\varepsilon)\begin{pmatrix} \xi_+ \\ \xi_- \end{pmatrix}+(H_\varepsilon h)(t), \qquad 0\le t\le 1, \tag{4.14}$$

where

$$\phi(t,\varepsilon)=\begin{bmatrix} \phi_+(t,\varepsilon) & 0 \\ 0 & \phi_-(t,\varepsilon) \end{bmatrix}, \tag{4.15}$$

$$(H_\varepsilon h)(t)=\begin{pmatrix} H_\varepsilon^+ h_+ \\ H_\varepsilon^- h_- \end{pmatrix}(t)=\begin{pmatrix} \frac{1}{\varepsilon}\int_1^t \phi_+(t,\varepsilon)\phi_+^{-1}(s,\varepsilon)h_+(s,\varepsilon)\,ds \\ \frac{1}{\varepsilon}\int_0^t \phi_-(t,\varepsilon)\phi_-^{-1}(s,\varepsilon)h_-(s,\varepsilon)\,ds \end{pmatrix} \tag{4.16}$$

have been set.

Proceeding as in §2 and using the estimates given in Lemma 4.1 yield

$$\|H_\varepsilon\|_{[0,1]} \leq c\frac{1}{\sqrt{\varepsilon}} \tag{4.17}$$

and

$$\|(H_\varepsilon f)(t)\| \leq \text{const}\,\min\left(\frac{1}{t}, \frac{1}{\sqrt{\varepsilon}}\right)\|f\|_{[0,1]}, \qquad t\in[0,1]. \tag{4.18}$$

We set

$$\bar{B}(t,\varepsilon) = B(t,\varepsilon)E(t) - E'(t) \tag{4.19}$$

and solve (4.8):

$$u(t,\varepsilon) = \phi(t,\varepsilon)\begin{pmatrix}\xi_+\\ \xi_-\end{pmatrix} + \varepsilon\big(H_\varepsilon E^{-1}(\cdot)\bar{B}(\cdot,\varepsilon)u\big)(t) + \big(H_\varepsilon E^{-1}(\cdot)f\big)(t). \tag{4.20}$$

Because of (4.17) the operator $I-\varepsilon H_\varepsilon E^{-1}\bar{B}$ is invertible on $C([0,1])$ and its inverse is uniformly bounded as $\varepsilon\to 0$. We get

$$u(t,\varepsilon) = \Omega(t,\varepsilon)\begin{pmatrix}\xi_+\\ \xi_-\end{pmatrix} + (G_\varepsilon f)(t), \qquad 0\leq t\leq 1, \tag{4.21}$$

where we denoted

$$\Omega(t,\varepsilon) = \Big(\big(I-\varepsilon H_\varepsilon E^{-1}\bar{B}\big)^{-1}\phi(\cdot,\varepsilon)\Big)(t), \tag{4.22a}$$

$$(G_\varepsilon f)(t) = \Big(\big(I-\varepsilon H_\varepsilon E^{-1}\bar{B}\big)^{-1}H_\varepsilon E^{-1}f\Big)(t). \tag{4.22b}$$

Using the series expansion of (4.22a, b) and the estimate in Lemma 4.1 we get

$$\Omega(t,\varepsilon) = \phi(t,\varepsilon) + O\left(t\exp\left(-\lambda_+\frac{t^2}{2\varepsilon}\right) + (1-t)\exp\left(\lambda_+\frac{t^2-1}{2\varepsilon}\right)\right), \qquad \varepsilon\to 0, \tag{4.23}$$

uniformly on [0, 1] and

$$\|(G_\varepsilon f)(t)\| \leq \text{const}\,\min\left(\frac{1}{t}, \frac{1}{\sqrt{\varepsilon}}\right)\|f\|_{[0,1]}. \tag{4.24}$$

Inserting into the boundary condition (4.2) gives the linear system of equations

$$\begin{aligned}&\big(B_0(\varepsilon)E(0)\Omega(0,\varepsilon) + B_1(\varepsilon)E(1)\Omega(1,\varepsilon)\big)\begin{pmatrix}\xi_+\\ \xi_-\end{pmatrix}\\ &\quad = \beta(\varepsilon) - B_0(\varepsilon)E(0)(G_\varepsilon f)(0) - B_1(\varepsilon)E(1)(G_\varepsilon f)(1).\end{aligned} \tag{4.25}$$

We partition

$$E(t) = [\underbrace{E_+(t)}_{r_+}, \underbrace{E_-(t)}_{r_-}] \tag{4.26}$$

and derive from (4.25), (4.23) that the linear system (4.25) is uniquely solvable if

$$[B_0(0)E_-(0), B_1(0)E_+(1)] \tag{4.27}$$

is nonsingular and $\xi_+(\varepsilon)$, $\xi_-(\varepsilon)$ fulfill (3.41). If $B_0(\varepsilon)E_+(0)\equiv 0$ or if f fulfills (3.32), (3.33) then $\|\xi_+(\varepsilon)\|$, $\|\xi_-(\varepsilon)\| \leq \text{const}(\|\beta\| + \|f\|_{[0,1]})$ holds.

Assume now that (4.27) holds and that f fufills (3.32), (3.33). Then the estimate (3.35) holds and if $g \in C^1([0,1])$ uniformly as $\varepsilon \to 0$ then $y(t,\varepsilon)$ fulfills (3.37) where $\bar{y}(t) = -A^{-1}(t)g(t,0)$ is the solution of the reduced equation (4.1).

For general f we get

THEOREM 4.1. *Assume that* (4.27) *holds, that* $f \in C([0,1])$ *uniformly in* ε *and that* f *fulfills* (3.39). *Then there is a unique solution* $y(t,\varepsilon)$ *of* (4.1), (4.2) (*for* ε *sufficiently small*) *which fulfills the estimate* (3.42). *If* $f \in C^1([0,1])$ *uniformly in* ε, *then* y *has the expansion given in Theorem* 3.2 (i) *where* $\bar{y}(t) = -A^{-1}(t)f(t,0)/t$ *has been set. If* $B_0(\varepsilon)E_+(0) \equiv 0$ *then the factor* $1/\sqrt{\varepsilon}$ *drops out. The estimate in Theorem* 3.2 (iii) *holds.*

So if the eigenvalues of $A(t)$ split into one group with strictly positive real parts and another group with strictly negative real parts (unstable and stable eigenvalues), then the results obtained for constant coefficient problems carry over. The solution $y(t,\varepsilon)$ has (generally) a boundary layer at $t=1$ of width $O(\varepsilon|\ln\varepsilon|)$ and height $O(1/\sqrt{\varepsilon})$ and decays exponentially (as $\varepsilon \to 0$) to the solution of the reduced problem (disregarding $O(\sqrt{\varepsilon})$ terms) within the layer at $t=1$.

Another boundary layer (of width $O(\sqrt{\varepsilon})$ and height $O(1/\sqrt{\varepsilon})$) occurs (generally) at $t=0$ (at the boundary turning point). However, this layer is of a different nature. The solution does not approach a smooth function exponentially, inside this layer. It converges to a possibly (around 0) unbounded function as $\varepsilon \to 0$.

5. Nonlinear problems with a boundary turning point. Using straightforward contraction arguments we can analyse quasilinear (see O'Malley (1978), Ringhofer (1981)) problems where the reduced solution has a first order pole at $t=0$:

$$\varepsilon y' = tA(t)y + \varepsilon h(y,t,\varepsilon) + f(t,\varepsilon), \qquad 0 \le t \le 1, \tag{5.1}$$

$$B_0(\varepsilon)y(0,\varepsilon) + B_1(\varepsilon)y(1,\varepsilon) = \beta(\varepsilon). \tag{5.2}$$

We assume that (4.3) holds, that $A(t)$ fulfills (4.4), (4.5), (4.6) and $f \in C([0,1])$ uniformly as $\varepsilon \to 0$.

Moreover we assume that the "semi"-reduced problem

$$\varepsilon z' = tA(t)z + \bar{h}(t,\varepsilon), \qquad 0 \le t \le 1, \tag{5.3}$$

$$B_0(\varepsilon)z(0,\varepsilon) + B_1(\varepsilon)z(1,\varepsilon) = \lambda(\varepsilon) \tag{5.4}$$

is stable, i.e., there is a unique solution z for all $\bar{h} \in C([0,1])$, $\lambda(\varepsilon) \in \mathbb{R}^n$, and the estimate

$$\|z(t)\| \le \text{const}\left(\|\lambda(\varepsilon)\| + \left(\frac{1}{\sqrt{\varepsilon}} \left(\exp\left(\lambda_+ \frac{t^2-1}{2\varepsilon} \right) + \exp\left(-\lambda_+ \frac{t^2}{2\varepsilon} \right) \right) + \min\left(\frac{1}{\sqrt{\varepsilon}}, \frac{1}{t} \right) \right) \|\bar{h}\|_{[0,1]} \right) \tag{5.5}$$

holds. Sufficient for this is that (4.27) holds.

The assumptions on h are

$$h \in C^1(\mathbb{R}^n) \cap C(\mathbb{R}^n \times [0,1]) \quad \text{uniformly for } t \in [0,1] \quad \text{as } \varepsilon \to 0 \tag{5.6}$$

and a growth restriction on $\partial h/\partial y$:

$$\left\| \frac{\partial h}{\partial y}(z,t,\varepsilon) \right\| \le \frac{c}{\sqrt{\varepsilon}} o(1) \quad \text{for } \|z\| \le \frac{c}{\sqrt{\varepsilon}} \quad \text{as } \varepsilon \to 0 \tag{5.7}$$

uniformly for $t \in [0,1]$.

We merely state

THEOREM 5. (i) *Under the given assumptions the problem* (5.1), (5.2) *has a solution* y *for* ε *sufficiently small.*

(ii) *If* $f\in C^1([0,1])$ *uniformly in* ε *and if* f *fulfills* (3.39), *then the expansion*

$$y(t,\varepsilon)=\bar{y}(t)+\frac{1}{\sqrt{\varepsilon}}\left(\sigma_-\left(\frac{t^2}{2\varepsilon}\right)+\sigma_+\left(\frac{1-t^2}{2\varepsilon}\right)\right)+o(1)+\frac{\varepsilon}{t^3}\rho(t,\varepsilon),\qquad \varepsilon\to 0, \tag{5.8}$$

holds for $0<t=t(\varepsilon)$ *and* $t(\varepsilon)/\sqrt{\varepsilon}\to\infty$. ρ *fulfills the estimate* (3.43) *and* $\bar{y}(t)=-A^{-1}(t)f(t,0)/t$ *solves the reduced problem* (5.1).

(iii) *If* $B_0(\varepsilon)E_+(0)\equiv 0$ *then the factor* $1/\sqrt{\varepsilon}$ *in* (5.8) *drops out.*

(iv) *Moreover,* y *is unique in* $K_{\rho,\varepsilon}=\{\bar{h}\in C([0,1])|\bar{h}(t,\varepsilon)=\bar{H}(t,\varepsilon)/\sqrt{\varepsilon},\ \|\bar{H}\|_{[0,1]}\le\rho\}$.

Now assume that $f(t,\varepsilon)=tg(t,\varepsilon)$, $g\in C^1([0,1])$, uniformly as $\varepsilon\to 0$. Then $\bar{y}(t)=-A^{-1}(t)g(t,0)\in L^\infty([0,1])$. In this case we can relax (5.7) by requiring

$$h\in C^2(\mathbb{R}^n)\cap C(\mathbb{R}^n\times[0,1])\quad\text{uniformly for } t\in[0,1]\quad\text{as } \varepsilon\to 0 \tag{5.9}$$

uniformly for $t\in[0,1]$. A statement analogous to Theorem 3.2 (ii) holds then.

Estimates on the derivatives of y for both cases of inhomogeneities can easily be derived by using the second part of Lemma 2.9.

Acknowledgment. Part of this research was done while the first author was visiting the Institut für Angewandte Mathematik of the Technische Universität Wien, Vienna, Austria.

REFERENCES

L. R. ABRAHAMSSON (1975), *A priori estimates for solutions of singular perturbations with a turning point*, Report 56, Dept. of Computer Sciences, Uppsala University, Uppsala, Sweden.

R. C. ACKERBERG AND R. E. O'MALLEY, JR. (1970), *Boundary layer problems exhibiting personance*, Stud. Appl. Math., 45, pp. 277–295.

U. ASCHER AND R. WEISS (1981), *Collocation for singular perturbation problems* I, *First order systems with constant coefficients*, TR 81–2, Univ. British Columbia, Vancouver, BC, Canada.

N. DUNFORD AND J. SCHWARTZ (1957), *Linear Operators*, Part 1, Interscience, New York.

F. HOWES (1980), *Some old and new results on singularly perturbed boundary value problems*, in Singular Perturbations and Asymptotics, Proceedings, R. E. Meyer and S. V. Parter, eds., Academic Press, New York.

H. O. KREISS AND S. V. PARTER (1974), *Remarks on singular perturbations with turning points*, this Journal, 5, pp. 230–251.

H. O. KREISS AND N. NICHOLS (1975), *Numerical methods for singular perturbation problems*, Report 57, Dept. of Computer Science, Uppsala University, Uppsala, Sweden.

P. A. MARKOWICH AND C. RINGHOFER (1983), *Singular perturbation problems with a singularity of the second kind*, this Journal, this issue, pp. 897–914.

R. E. O'MALLEY (1970), *On boundary value problems for a singularly perturbed differential equation with a turning point*, this Journal, 1, pp. 479–490.

R. E. O'MALLEY, JR. (1978), *On singular singularly perturbed initial value problems*, Applicable Anal., 8, pp. 79–81.

_____ (1979), *A singular singularly perturbed linear boundary value problem*, this Journal, 10, pp. 695–703.

C. A. RINGHOFER (1980), *Collocation methods for singularly perturbed boundary value problems*, Master's thesis, Technische Universität Wien, Vienna, Austria.

_____ (1981a), *A class of collocation schemes for singularly perturbed boundary value problems*, Thesis, Technische Universität Wien, Vienna, Austria.

_____ (1981b), *Existence and structure of solutions of nonlinear singularly perturbed boundary value problems*, manuscript.

H. L. TURRITIN (1952), *Asymptotic expansions of solutions of systems of ordinary differential equations containing a parameter*, Contrib. Theory Nonlinear Oscillations, 2, pp. 81–116.

W. WASOW (1978), *Topics in the theory of linear ordinary differential equations having singularities with respect to a parameter*, Rep. Institut de Recherche Mathematique Avances, Université Louis Pasteur, Strasbourg, France.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0005 $01.25/0

SINGULAR PERTURBATION PROBLEMS WITH A SINGULARITY OF THE SECOND KIND*

PETER A. MARKOWICH† AND C. A. RINGHOFER‡

Abstract. This paper deals with singularly perturbed systems of ordinary differential equations posed as boundary value problems on an infinite interval. The system is assumed to consist of singularly perturbed equations and unperturbed equations and to have a singularity of the second kind at ∞. Under the assumption that there is no turning point, we derive uniform asymptotic expansions (as the perturbation parameter tends to zero) for the fast (perturbed) and slow (unperturbed) components uniformly on the whole infinite line. The second goal of the paper is to derive convergence estimates for the solutions of 'finite' singular perturbation problems obtained by cutting the infinite interval at a finite (far out) point and by substituting appropriate additional (asymptotic) boundary conditions at the far end. Using suitably chosen asymptotic boundary conditions the order of convergence is shown to depend only on the decay property of the 'finite' solution.

AMS-MOS subject classification (1980). Primary 34B15, 34C05, 34C11, 34D15, 34E05, 34A45

Key words. nonlinear boundary value problems, singular points, boundedness of solutions, singular perturbations, asymptotic expansion, theoretical approximation of solutions

1. Introduction.

In this paper we deal with the singular perturbation problem

$$\varepsilon y' = t^{\alpha} h(y,z,t,\varepsilon), \tag{1.1}$$

$$z' = t^{\alpha} g(y,z,t,\varepsilon), \qquad \alpha > -1, \quad t \in [1,\infty)^{-}, \tag{1.2}$$

$$\begin{pmatrix} y \\ z \end{pmatrix} \in C([1,\infty]), \tag{1.3}$$

$$F(\varepsilon)\begin{pmatrix} y(1,\varepsilon) \\ z(1,\varepsilon) \end{pmatrix} = \beta(\varepsilon), \tag{1.4}$$

where $0<\varepsilon \ll 1$, y, h are (real) n-vectors, z, g are (real) m-vectors. $F(\varepsilon)$ is a (real) $k\times(n+m)$-matrix, $\beta(\varepsilon)$ a (real) k-vector (the relationship between $n+m$ and k will be explained later) and $C([1,\infty])$ denotes the space of functions in $C([1,\infty))$ which have a finite limit as $t\to\infty$. For the solution $\binom{y}{z}$ of this problem we call y its fast component and z its slow component. The system (1.1), (1.2) has a singularity of the second kind at $t=\infty$ for $\alpha>-1$.Problems of this kind frequently occur in fluid mechanics, especially in boundary layer theory (see for example Schlichting (1959) for the Orr–Sommerfeld problem and Lagerstrom (1961) for a model of flow past an obstacle) whenever flows of high Reynolds number ($R\sim 1/\varepsilon$) over infinite media are investigated. Other applications occur in thermodynamics (see Lagerstrom and Casten (1972)).

Our assumptions on the problems (1.1), (1.2) are the following. We assume that $h,g\in C^2(\mathbb{R}^{n+m}\times[1,\infty]\times[0,\varepsilon_0])$ where ε_0 is sufficiently small, and that the system is quasilinear in the sense of Ringhofer (1981), which means that

$$h(y,z,t,\varepsilon) = A(z,t)y + f(z,y,t,\varepsilon) \tag{1.5}$$

* Received by the editors October 12, 1981, and in revised form June 26, 1982.

†Institut für Numerische und Angewandte Mathematik, Technische Universität Wien, Gusshausstrasse 27-29, A-1040 Vienna, Austria. The work of this author was sponsored by the U. S. Army under contract DAAG29-80-C-0041 and the Austrian Ministry for Science and Research. This material is based upon work supported by the National Science Foundation under grant MCS-7927062.

‡Institut für Numerische und Angewandte Mathematik, Technische Universität Wien, Gusshausstrasse 27-29, A-1040 Vienna, Austria. The work of this author was supported by the Österreichischen Fond zur Förderung von Wissenschaft und Forschung.

and

$$\frac{\partial f}{\partial y}=O(\varepsilon) \tag{1.6}$$

for $t\in[1,\infty]$, $\varepsilon\in[0,\varepsilon_0]$ and y, z in compact subsets of $\mathbb{R}^{n+m}$. Here $A(z,t)$ is an $n\times n$-matrix which is in block diagonal form

$$A(z,t)=\begin{bmatrix} A_+(z,t) & 0 \\ \underbrace{0}_{r_+} & \underbrace{A_-(z,t)}_{r_-} \end{bmatrix}\begin{matrix}\}r_+\\ \}r_-\end{matrix} \tag{1.7}$$

such that the real parts of the eigenvalues of $A_+(z,t)$ are strictly positive and the real parts of the eigenvalues of $A_-(z,t)$ are strictly negative for $t\in[1,\infty]$ and z in a compact subset of $\mathbb{R}^m$ in which the slow solution component remains for $t\geq 1$. So we exclude turning-point problems where A has one or more eigenvalues whose real parts change sign.

The first goal of our analysis is to study the asymptotic behavior of the solutions of (1.1), (1.2), (1.3) as $\varepsilon\to 0^+$ globally on $[1,\infty]$ and to find conditions on $F(\varepsilon)$ which guarantee the locally unique solvability of the singular boundary value problem (1.1), (1.2), (1.3), (1.4).

For this we use techniques already developed for "finite" singular perturbation problems such as for example matched asymptotic expansions (see O'Malley (1978), (1979), Ringhofer (1980), (1981)) and the theory of singular boundary value problems (see de Hoog and Weiss (1980a, b), Markowich (1982a, b), (1983) and Lentini and Keller (1980)).

We show that there are solutions y, z of (1.1), (1.2), (1.3), (1.4) which have the form

$$z(t,\varepsilon)=\bar{z}(t)+O(\varepsilon), \qquad t\in[1,\infty], \tag{1.8}$$

$$y(t,\varepsilon)=\sigma\left(\frac{t-1}{\varepsilon}\right)+\bar{y}(t)+O(\varepsilon), \qquad t\in[1,\infty], \tag{1.9}$$

where $\bar{y}$, $\bar{z}$ are solutions of the reduced infinite problem, obtained by setting $\varepsilon=0$ in (1.1), (1.2) and (1.3), which satisfy appropriate boundary conditions. Here $\sigma(\tau)$ decays exponentially to zero as $\tau\to\infty$ (boundary layer term) and $\bar{z}$, $\bar{y}$ decay to a finite limit $\bar{z}_\infty$, $\bar{y}_\infty$ as $t\to\infty$ satisfying the equations

$$0=h(\bar{y}_\infty,\bar{z}_\infty,\infty,0), \tag{1.10}$$

$$0=g(\bar{y}_\infty,\bar{z}_\infty,\infty,0). \tag{1.11}$$

This result generalizes the results by O'Malley (1979) and Ringhofer (1980), (1981) obtained for finite interval singular perturbation problems.

Singularly perturbed initial value problems on the infinite line have been investigated by Hoppenstaedt (1966) under more stringent stability conditions than those considered here.

The second goal is to study approximating "finite" singular perturbation problems, which are set up by cutting the infinite interval $[1,\infty]$ at a finite point $T\gg 1$ and by substituting (for the continuity condition (1.3) at $t=\infty$) additional, so called asymptotic boundary conditions obtaining a "finite" singular perturbation problem

$$\varepsilon y_T'=t^\alpha h(y_T,z_T,t,\varepsilon), \tag{1.12}$$

$$z_T'=t^\alpha g(y_T,z_T,t,\varepsilon), \qquad 1\leq t\leq T, \tag{1.13}$$

$$F(\varepsilon)\begin{pmatrix} y_T(1,\varepsilon) \\ z_T(1,\varepsilon) \end{pmatrix} = \beta(\varepsilon), \tag{1.14}$$

$$S(T,\varepsilon)\begin{pmatrix} y_T(T,\varepsilon) \\ z_T(T,\varepsilon) \end{pmatrix} = \gamma(T,\varepsilon) \tag{1.15}$$

where $S(T,\varepsilon)$ is an $(n+m-k)\times(n+m)$-matrix, $\gamma(T,\varepsilon)\in\mathbb{R}^{n+m-k}$. The condition (1.15) shall reflect the asymptotic behavior of the "finite" solution (y,z) as $t\to\infty$.

Finite approximating two-point boundary value problems (for unperturbed infinite problems) have been studied extensively by de Hoog and Weiss (1980a), Markowich (1982b) and Lentini and Keller (1980).

We show that under rather mild assumptions on the "infinite" problem (a certain "wellposedness" is required), there is a choice of $S(T,\varepsilon)\equiv S$ and $\gamma(T,\varepsilon)\equiv\gamma$ only depending on the reduced infinite problem such that the "finite" (perturbed) problem has a unique solution y_T, z_T for T sufficiently large and ε sufficiently small (but T and ε independent) which fulfills the convergence estimate

$$\left\| \begin{pmatrix} y-y_T \\ z-z_T \end{pmatrix} \right\|_{[1,T]} \le K\left(\exp\left(-\frac{c}{\alpha+1}T^{\alpha+1}\right)+\varepsilon\right), \qquad K,c>0, \tag{1.16}$$

($\|\cdot\|_{[a,b]}$ denotes the sup-norm on $[a,b]$) where the constants K, c may be chosen independently of T and ε.

The "finite" singular perturbation problem (1.12), (1.13), (1.14), (1.15) can then be solved by polynomial collocation methods (see Kreiss and Nichols (1975), Ringhofer (1981), Ascher and Weiss (1981)). An exponential mesh size strategy for "long interval" problems has been developed for the box-scheme by Markowich and Ringhofer (1981). This can be used on $[\omega,T]$, $\omega>1$, while within the boundary layer (on $[1,1+O(\varepsilon|\ln\varepsilon|)]$) a very fine grid (see Ascher and Weiss (1981)) has to be used. Since the solution of (1.13), (1.14), (1.15), (1.16) is smooth (has derivatives which are uniformly bounded in ε) on $[1+O(\varepsilon|\ln\varepsilon|),\omega]$, standard techniques can be used there.

The paper is organized as follows. In §2 we prove estimates on the solution operators of perturbed and unperturbed linear constant coefficient problems on $[1,\infty]$; in §3 linear problems are dealt with, and §4 is concerned with nonlinear singular perturbation problems of the form (1.1)–(1.4).

2. Linear constant coefficient problems on infinite intervals. In this section we prove estimates on the solution operator of linear constant coefficient problems. These estimates will be needed for the variable coefficient case in §3.

First we investigate unperturbed problems of the form

$$z' = t^{\alpha}\tilde{J}z + t^{\alpha}g(t), \qquad \delta\le t<\infty, \quad \alpha>-1, \quad \delta\ge 1, \tag{2.1a}$$

$$z\in C([\delta,\infty]) \tag{2.1b}$$

where z, g are m-vectors and the $m\times m$-matrix $\tilde{J}$ is in Jordan canonical form

$$\tilde{J} = \begin{bmatrix} \tilde{J}^{+} & 0 \\ 0 & \tilde{J}^{-} \end{bmatrix} \begin{matrix} \}\tilde{r}_{+} \\ \}\tilde{r}_{-} \end{matrix} \tag{2.2}$$

$\tilde{J}^{+}$ has only eigenvalues with positive real parts and $\tilde{J}^{-}$ has only eigenvalues with negative real parts.

We define the fundamental matrix

$$\phi(t,\delta)=\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\alpha+1}\tilde{J}\right) \tag{2.3}$$

and the solution operator

$$(H_\delta g)(t)=\int_\infty^t \phi(t,\delta)\tilde{D}_+\phi^{-1}(s,\delta)s^\alpha g(s)\,ds+\int_\delta^t \phi(t,\delta)\tilde{D}_-\phi^{-1}(s,\delta)s^\alpha g(s)\,ds \tag{2.4}$$

where $\tilde{D}_+$, $\tilde{D}_-$ are diagonal projections

$$\tilde{D}_+=\begin{bmatrix} I_{\tilde{r}_+} & 0 \\ \underbrace{0}_{\tilde{r}_+} & \underbrace{0}_{\tilde{r}_-} \end{bmatrix}\begin{matrix}\}\tilde{r}_+\\ \}\tilde{r}_-\end{matrix} \tag{2.5a}$$

$$\tilde{D}_-=\begin{bmatrix} 0 & 0 \\ \underbrace{0}_{\tilde{r}_+} & \underbrace{I_{\tilde{r}_-}}_{\tilde{r}_-} \end{bmatrix}\begin{matrix}\}\tilde{r}_+\\ \}\tilde{r}_-\end{matrix} \tag{2.5b}$$

The general solution of (2.1) is given by

$$z(t)=\phi(t,\delta)\begin{bmatrix} 0 \\ I_{\tilde{r}_-} \end{bmatrix}\xi+(H_\delta g)(t), \qquad \xi\in\mathbb{C}^{\tilde{r}_-}, \tag{2.6}$$

if $g\in C([\delta,\infty])$. The operator H_δ has been extensively analyzed by de Hoog and Weiss (1980a, b) and Markowich (1982a, b), (1983). Therefore we only state its properties:

LEMMA 2.1. *Assume that $\tilde{J}$ fulfills* (2.2). *Then H_δ maps $C([\delta,\infty])$ into $C([\delta,\infty])$, its norm is bounded independently of δ and*

$$(H_\delta g)(\infty)=-\tilde{J}^{-1}g(\infty) \quad \textit{holds}. \tag{2.7}$$

We also denote the operator norm of $H_\delta: C([\delta,\infty])\to C([\delta,\infty])$ by $\|H_\delta\|_{[\delta,\infty]}$.

We now investigate singularly perturbed problems of the form

$$\varepsilon y'=t^\alpha Jy+t^\alpha f(t), \qquad t\geq\delta, \quad \alpha>-1, \tag{2.8a}$$

$$y\in C([\delta,\infty]) \tag{2.8b}$$

where y, f are n-vectors and J is in Jordan canonical form

$$J=\begin{bmatrix} J^+ & 0 \\ \underbrace{0}_{r_+} & \underbrace{J^-}_{r_-} \end{bmatrix}\begin{matrix}\}r_+\\ \}r_-\end{matrix} \tag{2.9}$$

such that the eigenvalues of J^+ have positive real parts and the eigenvalues of J^- have negative real parts. We define the fundamental matrix

$$\psi(t,\delta,\varepsilon)=\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\varepsilon(\alpha+1)}J\right) \tag{2.10}$$

and the solution operator

$$\begin{aligned}(G_{\varepsilon,\delta}f)(t)=&\frac{1}{\varepsilon}\int_\infty^t \psi(t,\delta,\varepsilon)D_+\psi^{-1}(s,\delta,\varepsilon)s^\alpha f(s)\,ds\\ &+\frac{1}{\varepsilon}\int_\delta^t \psi(t,\delta,\varepsilon)D_-\psi^{-1}(s,\delta,\varepsilon)s^\alpha f(s)\,ds\end{aligned} \tag{2.11}$$

where the diagonal projections D_+, D_- are defined analogously to (2.5). As in the unperturbed case, the general solution of (2.8) is given by

$$y(t,\varepsilon)=\psi(t,\delta,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\eta+(G_{\varepsilon,\delta}f)(t),\qquad \eta\in\mathbb{C}^{r_-}. \tag{2.12}$$

We prove

LEMMA 2.2. *Let J fulfill* (2.9). *Then $G_{\varepsilon,\delta}: C([\delta,\infty])\to C([\delta,\infty])$ is bounded independently of $\delta\geq 1$, $\varepsilon\in[0,\varepsilon_0]$. If $f, f'(t)/t^\alpha\in C([1,\infty])$ then*

$$(G_{\varepsilon,\delta}f)(t)=-J^{-1}f(t)+\psi(t,\delta,\varepsilon)D_-J^{-1}f(\delta)+\varepsilon(\Delta_{\varepsilon,\delta}f)(t) \tag{2.13}$$

holds, where

$$\|\Delta_{\varepsilon,\delta}f\|_{[\delta,\infty]}\leq K\max_{s\in[\delta,\infty]}\left\|\frac{f'(s)}{s^\alpha}\right\| \tag{2.14}$$

with K independent of ε, δ.

Proof. Obviously $(G_{\varepsilon,\delta}f)(t)$ can be written as

$$(G_{\varepsilon,\delta}f)(t)=\begin{pmatrix}\frac{1}{\varepsilon}\int_\infty^t\psi_+(t,\delta,\varepsilon)\psi_+^{-1}(s,\delta,\varepsilon)s^\alpha f_+(s)\,ds\\ \frac{1}{\varepsilon}\int_\delta^t\psi_-(t,\delta,\varepsilon)\psi_-^{-1}(s,\delta,\varepsilon)s^\alpha f_-(s)\,ds\end{pmatrix}=\begin{pmatrix}(G^+_{\varepsilon,\delta}f_+)(t)\\ (G^-_{\varepsilon,\delta}f_-)(t)\end{pmatrix} \tag{2.15}$$

where

$$f=\begin{pmatrix}f_+\\ f_-\end{pmatrix}\begin{matrix}\}r_+\\ \}r_-\end{matrix}$$

and

$$\psi_+(t,\delta,\varepsilon)=\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\varepsilon(\alpha+1)}J^+\right), \tag{2.16a}$$

$$\psi_-(t,\delta,\varepsilon)=\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\varepsilon(\alpha+1)}J^-\right) \tag{2.16b}$$

has been set. Applying the well-known representation of a holomorphic matrix function

$$\varphi(P)=\frac{1}{2\pi i}\int_{\Gamma_P}\varphi(\lambda)(\lambda I-P)^{-1}k\lambda, \tag{2.17}$$

where $\varphi:\Omega\subset\mathbb{C}\to\mathbb{C}$ holomorphically and P is a square matrix whose eigenvalues are enclosed by $\Gamma_P\subset\Omega$, gives

$$(G^+_{\varepsilon,\delta}f_+)(t)=\frac{1}{2\pi i}\int_{\Gamma_+}\left(\frac{1}{\varepsilon}\int_\infty^t\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\varepsilon(\alpha+1)}\lambda\right)s^\alpha(\lambda I-J^+)^{-1}f_+(s)\,ds\right)d\lambda, \tag{2.17a}$$

$$(G^-_{\varepsilon,\delta}f_-)(t)=\frac{1}{2\pi i}\int_{\Gamma_-}\left(\frac{1}{\varepsilon}\int_\delta^t\exp\left(\frac{t^{\alpha+1}-\delta^{\alpha+1}}{\varepsilon(\alpha+1)}\mu\right)s^\alpha(\mu I-J^-)^{-1}f_-(s)\,ds\right)d\mu \tag{2.17b}$$

where $\Gamma_+ \subset \{z \in \mathbb{C} | \mathrm{Re} z > 0\}$ encloses all eigenvalues of J^+ and $\Gamma_- \subset \{z \in \mathbb{C} | \mathrm{Re} z < 0\}$ encloses all eigenvalues of J^-. We get

$$(2.18)\quad \|(G_{\varepsilon,\delta} g)(t)\| \leq \mathrm{const}\left(\max_{\lambda \in \Gamma_+} \frac{1}{\varepsilon} \int_t^\infty \exp\left(\frac{t^{\alpha+1} - s^{\alpha+1}}{\varepsilon(\alpha+1)} \mathrm{Re}\lambda \right) s^\alpha ds \right.$$
$$\left. + \max_{\mu \in \Gamma_-} \frac{1}{\varepsilon} \int_\delta^t \exp\left(\frac{t^{\alpha+1} - s^{\alpha+1}}{\varepsilon(\alpha+1)} \mathrm{Re}\mu \right) s^\alpha ds \right) \| f \|_{[\delta,\infty]}.$$

Since $\mathrm{Re}\lambda \geq c_+ > 0$, $\mathrm{Re}\mu \leq -c_- < 0$ we immediately obtain

$$(2.19)\qquad \| G_{\varepsilon,\delta} \|_{[\delta,\infty]} \leq K$$

where K is independent of δ and ε.

Applying integration by parts to the inner integrals in (2.7) gives

(2.20a)

$$(G^+_{\varepsilon,\delta} f_+)(t) = -\frac{1}{2\pi i} \int_{\Gamma_+} \frac{1}{\lambda} (\lambda I - J^+)^{-1} d\lambda f_+(t)$$
$$+ \frac{\varepsilon}{2\pi i} \int_{\Gamma_+} \frac{1}{\lambda \varepsilon} \int_\infty^t \exp\left(\frac{t^{\alpha+1} - s^{\alpha+1}}{\varepsilon(\alpha+1)} \lambda \right) s^\alpha (\lambda I - J^+) \frac{f'_+(s)}{s^\alpha} ds\, d\lambda,$$

$$(2.20b)\quad (G^-_{\varepsilon,\delta} f_-)(t) = -\frac{1}{2\pi i} \int_{\Gamma_-} \frac{1}{\mu} (\mu I - J^-)^{-1} d\mu f_-(t)$$
$$- \frac{1}{2\pi i} \int_{\Gamma_-} \frac{1}{\mu} \exp\left(\frac{t^{\alpha+1} - \delta^{\alpha+1}}{\varepsilon(\alpha+1)} \mu \right) (\mu I - J)^{-1} d\mu f_-(\delta)$$
$$+ \frac{\varepsilon}{2\pi i} \int_{\Gamma_-} \frac{1}{\mu \varepsilon} \int_\delta^t \exp\left(\frac{t^{\alpha+1} - s^{\alpha+1}}{\varepsilon(\alpha+1)} \mu \right) (\mu I - J^-)^{-1} \frac{f'_-(s)}{s^\alpha} ds\, d\mu.$$

Using (2.17) for the first term on the right-hand side of (2.20a) and for the first two terms on the right-hand side of (2.20b) and applying (2.19) to the last term on the right-hand sides of (2.20a, b) gives (2.13).

The term $-J^{-1} f(t)$ is the solution of the reduced problem (2.8), obtained by setting $\varepsilon = 0$, and

$$\psi(t,\delta,\varepsilon) \begin{bmatrix} 0 \\ I_{r_-} \end{bmatrix} \left(\eta + (J^-)^{-1} f_-(\delta) \right)$$

is the boundary layer term which decays as $\exp(-c(t^{\alpha+1} - \delta^{\alpha+1})/\varepsilon(\alpha+1))$, $c > 0$ as $\varepsilon \to 0+$. The boundary layer is located in $[\delta, \delta + O(\varepsilon |\ln \varepsilon|)]$. Lemma 2.2 generalizes the analogous result for finite interval singular perturbation problems.

In the sequel we will need the space

$$(2.21)\quad C^1_\alpha([\delta,\infty]) := C([\delta,\infty]) \cap C^1([\delta,\infty]) \cap \left\{ f \,\middle|\, \max_{s \in [\delta,\infty]} \left\| \frac{f'(s)}{s^\alpha} \right\| < \infty \right\},$$

which is equipped with the norm

$$(2.22)\qquad \| f \|^\alpha_{[\delta,\infty]} = \| f \|_{[\delta,\infty]} + \max_{s \in [\delta,\infty]} \left\| \frac{f'(s)}{s^\alpha} \right\|.$$

Then $H_\delta: C([\delta,\infty])\to C^1_\alpha([\delta,\infty])$, $\Delta_{\varepsilon,\delta}: C^1_\alpha([\delta,\infty])\to C([\delta,\infty])$ and

$$\|H_\delta\|_{C([\delta,\infty])\to C^1_\alpha([\delta,\infty])}\le K_1, \tag{2.23}$$

$$\|\Delta_{\varepsilon,\delta}\|_{C^1_\alpha([\delta,\infty])\to C([\delta,\infty])}\le K_2 \tag{2.24}$$

where K_1, K_2 are independent of δ, ε.

3. Variable coefficient problems. We consider the problem

$$\varepsilon y' = t^\alpha A(t,\varepsilon)y + t^\alpha B(t,\varepsilon)z + t^\alpha f(t,\varepsilon), \tag{3.1}$$

$$z' = t^\alpha C(t,\varepsilon)y + t^\alpha D(t,\varepsilon)z + t^\alpha g(t,\varepsilon), \qquad 1\le t<\infty,\quad \alpha>-1, \tag{3.2}$$

$$F(\varepsilon)\begin{pmatrix} y(1,\varepsilon)\\ z(1,\varepsilon)\end{pmatrix} = \beta(\varepsilon), \tag{3.3}$$

$$\begin{pmatrix} y\\ z\end{pmatrix}\in C([1,\infty]) \tag{3.4}$$

where the dimensions are as in §1 and assume that

$$A,B,C,D,f,g\in C([1,\infty]\times[0,\varepsilon_0]),\qquad F,\beta\in C([0,\varepsilon_0]), \tag{3.5a}$$

$$A,B,f\in C^1_\alpha([1,\infty]) \quad \text{uniformly in } \varepsilon\in[0,\varepsilon_0] \tag{3.5b}$$

holds for some $\varepsilon_0>0$ and that F, β, A, B, C, D, f, g are uniformly Lipschitz continuous at $\varepsilon=0$.

Moreover we assume that the eigenvalues $\lambda(t)$ of $A(t,0)$ split up into two groups such that

$$\mathrm{Re}\lambda_1(t)\ge c_+,\cdots,\mathrm{Re}\lambda_{r_+}(t)\ge c_+,\qquad c_+>0,\quad t\ge 1, \tag{3.6}$$

$$\mathrm{Re}\lambda_{r_++1}(t)\le -c_-,\cdots,\mathrm{Re}\lambda_n(t)\le -c_-,\qquad c_->0,\quad t\ge 1\quad (n-r_+=r_-) \tag{3.7}$$

holds (eigenvalues are counted according to algebraic multiplicities) and that there is a transformation to block form

$$A(t,0)=E(t)J(t)E^{-1}(t),\qquad J(t)=\left[\begin{array}{c|c} J_+(t) & 0\\ \hline 0 & J_-(t)\end{array}\right]\begin{matrix}\}r_+\\ \}r_-\end{matrix},\qquad t\ge 1, \tag{3.8}$$

such that the eigenvalues of $J_+(t)$ ($J_-(t)$) are $\lambda_1(t),\cdots,\lambda_{r_+}(t)$ ($\lambda_{r_++1}(t),\cdots,\lambda_n(t)$) and

$$E,E^{-1}\in C^1_\alpha([1,\infty]). \tag{3.9}$$

J is not necessarily in Jordan canonical form; however, we choose E such that $J(\infty)$ is in Jordan form. Under the assumptions (3.6), (3.7) and additional smoothness assumptions on A, a transformation matrix $E(t)$ exists for every $t\in[1,\infty]$, but the assumption (3.9) is much more restrictive (see O'Malley (1979)). At first we investigate

$$\varepsilon y' = t^\alpha A(t,\varepsilon)y + t^\alpha h(t,\varepsilon),\qquad y\in C([1,\infty]). \tag{3.10}$$

LEMMA 3.1. *Let A fulfill* (3.5)–(3.9) *and assume that $h\in C([1,\infty]\times[0,\varepsilon_0])$ and $h\in C^1_\alpha([1,\infty])$ uniformly for $\varepsilon\in[0,\varepsilon_0]$. Then the general solution of* (3.10) *satisfies*

$$y(t,\varepsilon)=\Lambda(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\Big(\zeta+\big[0,I_{r_-}\big]E^{-1}(1)A^{-1}(1,0)h(1,\varepsilon)\Big) \tag{3.11a}$$

$$-A^{-1}(t,0)h(t,\varepsilon)+O\big(\varepsilon\|h(\cdot,\varepsilon)\|^\alpha_{[1,\infty]}\big)$$

uniformly for $t\in[1,\infty]$ *with* $\zeta\in\mathbb{C}^r$. *The* $n\times n$*-matrix* $\Lambda(t,\varepsilon)$ *fulfills the estimate*

$$\left\|\Lambda(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\right\|\leq K_1\exp\left(-\frac{K_2}{(\alpha+1)\varepsilon}(t^{\alpha+1}-1)\right),\qquad t\geq 1, \tag{3.11b}$$

where $K_1, K_2>0$ *are independent of* t, ε *and*

$$\Lambda(1,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}=E(1)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}+O(\varepsilon).$$

Proof. We substitute

$$y=E(t)x \tag{3.12}$$

and obtain

$$\varepsilon x'=t^\alpha J(t)x+t^\alpha\big(E^{-1}(t)(A(t,\varepsilon)-A(t,0))E(t)-\varepsilon t^{-\alpha}E^{-1}(t)E'(t)\big)x + t^\alpha E^{-1}(t)h(t,\varepsilon), \tag{3.13}$$

$$x\in C([1,\infty]). \tag{3.14}$$

Using a perturbation approach we first solve

$$\varepsilon u'=t^\alpha J(t)u+t^\alpha d(t,\varepsilon),\qquad u\in C([1,\infty]), \tag{3.15}$$

where d fulfills the assumption on h stated in the lemma. According to (3.8) the system (3.15) splits up into

$$\varepsilon u'_+=t^\alpha J_+(t)u_++t^\alpha d_+(t,\varepsilon),\qquad u_+\in C([1,\infty]), \tag{3.16a}$$

$$\varepsilon u'_-=t^\alpha J_-(t)u_-+t^\alpha d_-(t,\varepsilon),\qquad u_-\in C([1,\infty]). \tag{3.16b}$$

At first we analyze (3.16a), which we rewrite as

$$\varepsilon u'_+=t^\alpha J_+(\infty)u_++t^\alpha\big(J_+(t)-J_+(\infty)\big)u_++t^\alpha d_+(t,\varepsilon), \tag{3.17a}$$

$$u_+\in C([1,\infty]). \tag{3.17b}$$

We regard (3.17) as an inhomogeneous constant coefficient problem with the fundamental matrix

$$\psi_+(t,\delta,\varepsilon)=\exp\left(\frac{J_+(\infty)}{\varepsilon(\alpha+1)}(t^{\alpha+1}-\delta^{\alpha+1})\right),\qquad \delta\geq 1, \tag{3.18}$$

and with solution operator

$$(G^+_{\varepsilon,\delta}d_+(\cdot,\varepsilon))(t)=\frac{1}{\varepsilon}\int_\infty^t\psi_+(t,\delta,\varepsilon)\psi_+^{-1}(s,\delta,\varepsilon)s^\alpha d_+(s,\varepsilon)\,ds,\qquad t\geq\delta \tag{3.19}$$

(as given in §1). Then, (3.17) can be rewritten as

$$u_+=G^+_{\varepsilon,\delta}\big(J_+(\cdot)-J_+(\infty)\big)u_++G^+_{\varepsilon,\delta}d_+(\cdot,\varepsilon). \tag{3.20}$$

Because of Lemma 2.2 and since $J_+(t)\to J_+(\infty)$, the operator $I-G^+_{\varepsilon,\delta}(J_+(\cdot)-J_+(\infty))$ is invertible on $C([\delta,\infty])$ for δ sufficiently large. We obtain

$$u_+(t)=\underbrace{\Big(\big(I-G^+_{\varepsilon,\delta}(J_+(\cdot)-J_+(\infty))\big)^{-1}G^+_{\varepsilon,\delta}d_+(\cdot,\varepsilon)\Big)(t)}_{(\theta^+_\varepsilon d_+(\cdot,\varepsilon))(t)},\qquad t\geq\delta. \tag{3.21}$$

To get a solution on $[1,\infty]$ we solve the terminal value problem

$$\varepsilon\tilde u'_+=t^\alpha J_+(t)\tilde u_++t^\alpha d_+(t,\varepsilon),\qquad 1\leq t\leq\delta, \tag{3.22}$$

$$\tilde u_+(\delta)=u_+(\delta) \tag{3.23}$$

and set $(\theta_\varepsilon^+ d_+(\cdot,\varepsilon))(t) := \tilde{u}_+(t)$ for $t \le \delta$. θ_ε^+ is an operator on $C([1,\infty])$ and since the eigenvalues of $J_+(t)$ have strictly positive real parts,

$$\|\theta_\varepsilon^+\|_{[1,\infty]} \le \text{const} \tag{3.24}$$

holds. Repeated application of (2.13), (2.14) gives

$$\begin{aligned}(\theta_\varepsilon^+ d_+(\cdot,\varepsilon))(t) &= \sum_{i=0}^{\infty} \left(\left(G_{\varepsilon,\delta}^+ (J_+(\cdot) - J_+(\infty))\right)^i G_{\varepsilon,\delta}^+ d_+(\cdot,\varepsilon)\right)(t) \\ &= \sum_{i=0}^{\infty} \left(\left(J_+(\infty)\right)^{-1} \left(J_+(t) - J_+(\infty)\right)\right)^i \left(J_+(\infty)\right)^{-1} d_+(t,\varepsilon) \\ &\quad + O\left(\varepsilon \| d_+(\cdot,\varepsilon)\|_{[\delta,\infty]}^{\alpha}\right) \\ &= -\left(J_+(t)\right)^{-1} d_+(t,\varepsilon) + O\left(\varepsilon \| d_+(\cdot,\varepsilon)\|_{[\delta,\infty]}^{\alpha}\right)\end{aligned} \tag{3.25}$$

for $t \ge \delta$. By continuation (3.25) holds for $t \ge 1$ (see Ringhofer (1981)).

We rewrite (3.16b) analogously

$$\varepsilon u_-' = t^\alpha J_-(\infty) u_- + t^\alpha \left(J_-(t) - J_-(\infty)\right) u_- + t^\alpha d_-(t,\varepsilon), \tag{3.26a}$$

$$u_- \in C([1,\infty]) \tag{3.26b}$$

and define the fundamental matrix

$$\psi_-(t,\delta,\varepsilon) = \exp\left(\frac{J_-(\infty)}{\varepsilon(\alpha+1)} \right)\left(t^{\alpha+1} - \delta^{\alpha+1}\right), \qquad \delta \ge 1, \tag{3.27}$$

and solution operator

$$\left(G_{\varepsilon,\delta}^- d_-(\cdot,\varepsilon)(t)\right) = \frac{1}{\varepsilon} \int_\delta^t \psi_-(t,\delta,\varepsilon) \psi_-^{-1}(s,\delta,\varepsilon) s^\alpha d_-(s,\varepsilon)\, ds, \qquad t \ge \delta, \tag{3.28}$$

such that the general solution of (3.26) is

$$\begin{aligned}u_- = &\left(I - G_{\varepsilon,\delta}^- \left(J_-(\cdot) - J_-(\infty)\right)\right)^{-1} \psi_-(\cdot,\delta,\varepsilon)\bar{\xi} \\ &+ \left(I - G_{\varepsilon,\delta}^- \left(J_-(\cdot) - J_-(\infty)\right)\right)^{-1} G_{\varepsilon,\delta} d_-(\cdot,\varepsilon)\end{aligned} \tag{3.29}$$

for $\bar{\xi} \in \mathbb{C}^{r_-}$ and $t \ge \delta$. We call the first term on the right-hand side of (3.29) $\bar{\psi}_-(t,\delta,\varepsilon)\bar{\xi}$ and the second $\bar{u}_{p_-}(t,\varepsilon)$. Obviously

$$\bar{\psi}_-(\delta,\delta,\varepsilon) = I_{r_-}, \quad \bar{\psi}_-(\infty,\delta,\varepsilon) = 0, \quad \bar{u}_{p_-}(\delta,\varepsilon) = 0 \tag{3.30}$$

hold. $\bar{\psi}_-$ has a boundary layer at δ. The homogeneous problem (3.16)(b) has a fundamental matrix $\tilde{\psi}_-(t,\varepsilon)$ such that

$$\tilde{\psi}_-(1,\varepsilon) = I_{r_-}, \tag{3.31a}$$

$$\|\tilde{\psi}_-(t,\varepsilon)\| \le c_1 \exp\left(\frac{-c_2}{\varepsilon(\alpha+1)} \left(t^{\alpha+1} - 1\right)\right) \tag{3.31b}$$

holds for $t \in [1,\delta]$ where $c_1, c_2 > 0$ (see Ringhofer (1981) and under more general assumptions O'Malley (1978)). We set

$$\check{\psi}_-(t,\varepsilon) = \bar{\psi}_-(t,\delta,\varepsilon)\tilde{\psi}_-(\delta,\varepsilon), \qquad t \ge 1. \tag{3.32}$$

Since $\check{\psi}_-(\delta,\varepsilon)=\tilde{\psi}_-(\delta,\varepsilon)$ we obtain $\check{\psi}_-\equiv\tilde{\psi}_-$ and the boundary layer has been shifted from δ to 1. On $[1+O(\varepsilon|\ln\varepsilon|),\infty]$ the matrix $\check{\psi}_-$ is smooth. Another particular solution is

$$\tilde{u}_{p_-}(t,\varepsilon)=\frac{1}{\varepsilon}\int_1^t \tilde{\psi}_-(t,\varepsilon)\check{\psi}_-^{-1}(s,\varepsilon)s^{\alpha}d_-(s,\varepsilon)\,ds, \qquad t\in[1,\delta]. \tag{3.33}$$

Since $\|\tilde{\psi}_-(t,\varepsilon)\tilde{\psi}_-^{-1}(s,\varepsilon)\|\le c_3\exp((-c_2/\varepsilon(\alpha+1))(t^{\alpha+1}-s^{\alpha+1}))$ holds on $[1,\delta]$ we derive

$$\|\tilde{u}_{p_-}(\cdot,\varepsilon)\|_{[1,\delta]}\le\text{const}\|d_+(\cdot,\varepsilon)\|_{[1,\delta]}. \tag{3.34}$$

Setting

$$\check{u}_{p_-}(t,\delta)=(\theta_\varepsilon^- d_-(\cdot,\varepsilon))(t):=\bar{\psi}_-(t,\delta,\varepsilon)\tilde{u}_{p_-}(\delta,\varepsilon)+\bar{u}_{p_-}(t,\varepsilon), \tag{3.35}$$

we obtain $\check{u}_{p_-}\equiv\tilde{u}_{p_-}$ and

$$\|\theta_\varepsilon^-\|_{[1,\infty]}\le\text{const} \tag{3.36}$$

because on $[1,\delta]$ we use (3.34) and on $[\delta,\infty]$ we use (3.35) and (3.29). As a general solution of (3.16b) we take

$$u_-(t,\varepsilon)=\check{\psi}_-(t,\varepsilon)\zeta+(\theta_\varepsilon^- d_-(\cdot,\varepsilon))(t), \qquad t\ge1, \tag{3.37}$$

and we get similarly to (3.25):

$$\begin{aligned}u_-(t,\varepsilon)=\check{\psi}_-(t,\varepsilon)\big(\zeta+J_-^{-1}(1)d_-(1,\varepsilon)\big)-J_-^{-1}(t)d_-(t,\varepsilon)\\ +O\big(\varepsilon\|d_-(\cdot,\varepsilon)\|_{[1,\infty]}^{\alpha}\big)\end{aligned} \tag{3.38}$$

uniformly on $[1,\infty]$.

Setting

$$\theta_\varepsilon=\begin{pmatrix}\theta_\varepsilon^+\\ \theta_\varepsilon^-\end{pmatrix}$$

we write the solution of (3.13), (3.14) as

$$x=\begin{bmatrix}0\\ \check{\psi}_-(\cdot,\varepsilon)\end{bmatrix}\zeta+\theta_\varepsilon\big(E^{-1}(A(\cdot,\varepsilon)-A(\cdot,0))E-\varepsilon\tilde{E}\big)x+\theta_\varepsilon E^{-1}h(\cdot,\varepsilon) \tag{3.39}$$

where $\tilde{E}(t)=t^{-\alpha}E^{-1}(t)E'(t)$ has been set. (3.5), (3.9) guarantee that $\tilde{A}(t,\varepsilon)=E^{-1}(t)(A(t,\varepsilon)-A(t,0))E-\varepsilon\tilde{E}(t)\to0$ as $\varepsilon\to0$ uniformly on $[1,\infty]$. From Lemma 2.2 we get that $(I-\theta_\varepsilon\tilde{A}(\cdot,\varepsilon))^{-1}$ exists on $C([1,\infty])$ for ε sufficiently small such that

$$\begin{aligned}x(t,\varepsilon)=\left(\big(I-\theta_\varepsilon\tilde{A}(\cdot,\varepsilon)\big)^{-1}\begin{bmatrix}0\\ \check{\psi}_-(\cdot,\varepsilon)\end{bmatrix}\right)(t)\zeta\\ +\Big(\big(I-\theta_\varepsilon\tilde{A}(\cdot,\varepsilon)\big)^{-1}\theta_\varepsilon E^{-1}h(\cdot,\varepsilon)\Big)(t), \qquad t\ge1,\end{aligned} \tag{3.40}$$

holds for $\zeta\in\mathbb{C}^{r_-}$.

(3.11a) follows from Lemma 2.2 and from $(I-\theta_\varepsilon\tilde{A}(\cdot,\varepsilon))^{-1}=I+O(\varepsilon)$ on $C([1,\infty])$.

$$\Lambda(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}=E(t)\begin{bmatrix}0\\ \check{\psi}_-(t,\varepsilon)\end{bmatrix}+O(\varepsilon)$$

is the boundary layer term (at $t=1$) fulfilling the estimate (3.31b) and $\zeta\in\mathbb{C}^{r_-}$.

Now we return to the coupled problem (3.1), (3.2), (3.3), (3.4). From Lemma 3.1 we get for fixed $z \in C_\alpha^1([1,\infty])$

$$y(t,\varepsilon)=\Lambda(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\rho - A^{-1}(t,0)\big(B(t,0)z(t,\varepsilon)+f(t,0)\big) + \varepsilon\big(L_\varepsilon^{(1)}z\big)(t)+\varepsilon\big(L_\varepsilon^{(2)}f\big)(t) \tag{3.41}$$

where $L_\varepsilon^{(i)} := C_\alpha^1([1,\infty]) \to C([1,\infty])$, $\|L_\varepsilon^{(i)}\|_{C_\alpha^1([1,\infty])\to C([1,\infty])} \le \text{const}$, $i=1, 2$, and $\rho = \zeta + [0, I_{r_-}]E^{-1}(1)A^{-1}(1,0)\big(B(1,0)z(1,\varepsilon)+f(1,0)\big)$. Inserting (3.41) into (3.2) gives

$$\begin{aligned} z' = {} & t^\alpha\big(D(t,0)-C(t,0)A^{-1}(t,0)B(t,0)\big)z + t^\alpha\varepsilon\big(L_\varepsilon^{(3)}z\big)(t) \\ & + t^\alpha\varepsilon\big(L_\varepsilon^{(4)}f\big)(t) + t^\alpha\Big(C(t,\varepsilon)\Lambda(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\rho \\ & + g(t,\varepsilon) - C(t,0)A^{-1}(t,0)f(t,0)\Big), \\ & z \in C([1,\infty]). \end{aligned} \tag{3.42}$$

Also the operators $L_\varepsilon^{(j)}: C_\alpha^1([1,\infty]) \to C([1,\infty])$, $j=3, 4$, are uniformly (in ε) bounded.

Setting $\tilde{D}(t,\varepsilon) = D(t,\varepsilon) - C(t,\varepsilon)A^{-1}(t,\varepsilon)B(t,\varepsilon)$ we have to solve a problem of the form:

$$z' = t^\alpha \tilde{D}(t,0)z + t^\alpha \tilde{g}(t,\varepsilon), \qquad z \in C_\alpha^1([1,\infty]), \tag{3.43}$$

where $\tilde{g} \in C([1,\infty]\times[0,\varepsilon_0])$. We apply Lemma 2.1 and the theory developed by de Hoog and Weiss (1980a, b). Therefore we assume that the Jordan form $\tilde{J}$ of $\tilde{D}(\infty,0)$, obtained by $\tilde{D}(\infty,0) = E_2 \tilde{J} E_2^{-1}$, has the block form given in (2.2). We obtain for $\xi \in \mathbb{C}^{\tilde{r}_-}$

$$\begin{aligned} z = {} & E_2\big(I - E_2 H_\delta E_2^{-1}\big(\tilde{D}(\cdot,0)-\tilde{D}(\infty,0)\big)E_2\big)^{-1}\phi(\cdot,\delta)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi \\ & + E_2\big(I - E_2 H_\delta E_2^{-1}\big(\tilde{D}(\cdot,0)-\tilde{D}(\infty,0)\big)E_2\big)^{-1} H_\delta E_2^{-1}\tilde{g}(\cdot,\varepsilon), \qquad t \ge \delta, \end{aligned} \tag{3.44}$$

where H_δ, $\phi(t,\delta)$ are defined in (2.24), (2.23) respectively and δ is sufficiently large. The right-hand side of (3.44) can be continued to $[1,\delta]$, and we obtain

$$z(t,\varepsilon) = \check{\phi}(t)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi + \big(\Gamma\tilde{g}(\cdot,\varepsilon)\big)(t), \qquad t \ge 1, \tag{3.45}$$

where $\Gamma: C([1,\infty]) \to C_\alpha^1([1,\infty])$ and $\|\Gamma\|_{C([1,\infty])\to C_\alpha^1([1,\infty])} \le \text{const}$ (see (2.23)). Applying this to (3.42) gives

$$\begin{aligned} z(t,\varepsilon) = {} & \check{\phi}(t)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi + \varepsilon\big(\Gamma L_\varepsilon^{(3)}z\big)(t) + \varepsilon\big(\Gamma L_\varepsilon^{(4)}f\big)(t) + \big(\Gamma g(\cdot,\varepsilon)\big)(t) \\ & + \left(\Gamma\left(C(\cdot,\varepsilon)\Lambda(\cdot,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\right)(t)\rho - \big(\Gamma C(\cdot,0)A^{-1}(\cdot,0)f(\cdot,0)\big)(t)\right), \end{aligned} \tag{3.46}$$

$\Gamma L_\varepsilon^{(3)}: C_\alpha^1([1,\infty]) \to C_\alpha^1([1,\infty])$ and $\|\Gamma L_\varepsilon^{(3)}\|_{[1,\infty]}^\alpha \le \text{const}$. Therefore $(I - \varepsilon\Gamma L_\varepsilon^{(3)})^{-1}$ exist as operators on $C_\alpha^1([1,\infty])$ for ε sufficiently small and

$$\begin{aligned} z(t,\varepsilon) = {} & \check{\phi}(t)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi + \big(\Gamma\big(g(\cdot,0)-C(\cdot,0)A^{-1}(\cdot,0)f(\cdot,0)\big)\big)(t) \\ & + \left(\Gamma C(\cdot,\varepsilon)\Lambda(\cdot,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\right)(t)\rho + O(\varepsilon). \end{aligned} \tag{3.47}$$

Using the exponential decay of $\Lambda(t,\varepsilon)[\begin{smallmatrix}0\\ I_{r_-}\end{smallmatrix}]$ and the definition of H_δ, it is easy to show that

$$\left\| \Gamma C(\cdot,\varepsilon)\Lambda(\cdot,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\right\|_{[1,\infty]} = O(\varepsilon). \tag{3.48}$$

So we obtain

THEOREM 3.1. *Assume that* (3.5)–(3.9) *hold and that the Jordan form of* $\tilde{D}(\infty,0)$ *fulfills* (2.2). *Also assume that the* $(r_-+\tilde{r}_-)\times(r_-+\tilde{r}_-)$*-matrix* $\hat{F}$ *defined by*

$$\hat{F}=F(0)\left[\begin{array}{c|c} E(1)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix} & E(1)D_-E^{-1}(1)A^{-1}(1,0)B(1,0)\check{\phi}(1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix} \\ \hline 0 & \check{\phi}(1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\end{array}\right] \tag{3.49}$$

where D_- *is defined in* §2, *is nonsingular* ($F(\varepsilon)$ *is an* $(r_-+\tilde{r}_-)\times(n+m)$*-matrix*). *Then the boundary value problem* (3.1), (3.2), (3.3), (3.4) *has for sufficiently small* ε *and for all* f *which fulfill* (3.5) *and for all* $r_-+\tilde{r}_-$*-vectors* β *a unique solution* y, z *which depends uniformly* (*in* ε) *continuously on* $f\in C^1_\alpha([1,\infty])$ *and* $g\in C([1,\infty])$ *when regarded as dwelling in* $C([1,\infty])$.

Moreover $y,z\in C^1([\delta,\infty])$ *for* $\delta>1$ *depends uniformly continuously on* $f\in C^1([1,\infty])$ *and* $g\in C([1,\infty])$ *and* β.

Proof. Inserting (3.47), (3.41) into the boundary condition (3.3) gives the system of linear equations

$$(\hat{F}+O(\varepsilon))\begin{pmatrix}\zeta\\ \xi\end{pmatrix}=\beta(\varepsilon)+\mathcal{F}_\varepsilon(f)+\mathcal{G}_\varepsilon(g) \tag{3.50}$$

where $\mathcal{F}_\varepsilon$, $\mathcal{G}_\varepsilon$ are uniformly bounded linear functionals from $C^1_\alpha([1,\infty])$ into $\mathbb{R}$ and $C([1,\infty])$ into $\mathbb{R}$ respectively.

The nonsingularity of $\hat{F}$ implies the solvability of (3.50) for ε sufficiently small.

A slight modification shows that the existence results of Theorem 3.1 also hold if only $f,g\in C([1,\infty])$ is assumed, and then $y,z\in C([1,\infty])$ depend uniformly continuously on $f,g\in C([1,\infty])$.

Theorem 3.1 enables us to define boundary conditions for the reduced problem. Therefore we partition $F(0)$:

$$F(0)=[\underbrace{F_y(0)}_{n},\underbrace{F_z(0)}_{m}]\}r_-+\tilde{r}_-. \tag{3.51}$$

Since $\hat{F}$ is nonsingular, the matrix

$$F_y(0)E(1)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}$$

has (full) rank r_-. Therefore, there is a nonsingular $(r_-+\tilde{r}_-)\times(r_-+\tilde{r}_-)$-matrix Y such that

$$YF_y(0)E(1)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}=\begin{bmatrix}U\\ \underbrace{0}_{\tilde{r}_-}\end{bmatrix}\begin{matrix}\}r_-\\ \}\tilde{r}_-\end{matrix} \tag{3.52}$$

holds where U is nonsingular. We partition Y by

$$\begin{matrix}\left[\begin{matrix}Y_1\\ Y_2\end{matrix}\right]\begin{matrix}\}r_-\\ \}\tilde{r}_-\end{matrix}\\ \underbrace{\qquad}_{r_-+\tilde{r}_-}\end{matrix} \tag{3.53}$$

The matrix

$$YF=\begin{bmatrix}U & U_1\\ 0 & U_2\end{bmatrix} \tag{3.54}$$

is block-upper triangular, and the variable ξ, which determines the slow component z, is the solution of $(U_2+O(\varepsilon))\xi+O(\varepsilon)\zeta=Y_2(\beta(\varepsilon)+\mathfrak{F}_\varepsilon(f)+\mathfrak{G}_\varepsilon(g))$. We therefore define the reduced problem as

$$0=A(t,0)\bar{y}(t)+B(t,0)\bar{z}(t)+f(t,0), \tag{3.55}$$

$$\bar{z}'=t^\alpha C(t,0)\bar{y}+t^\alpha D(t,0)\bar{z}+t^\alpha g(t,0), \qquad 1\le t<\infty, \tag{3.56}$$

$$Y_2 f(0)\begin{pmatrix}\bar{y}(1)\\ \bar{z}(1)\end{pmatrix}=Y_2\beta(0), \tag{3.57}$$

$$\begin{pmatrix}\bar{y}\\ \bar{z}\end{pmatrix}\in C([1,\infty]) \tag{3.58}$$

and get under the assumptions of Theorem 3.1

$$y(t,\varepsilon)=\bar{y}(t)+\sigma\left(\frac{t}{\varepsilon}\right)+O(\varepsilon), \tag{3.59}$$

$$z(t,\varepsilon)=\bar{z}(t)+O(\varepsilon) \tag{3.60}$$

uniformly on $[1,\infty]$ where $\sigma(t/\varepsilon)$ is the layer term which satisfies the estimate (3.11)(b). (3.59), (3.60) follow from (3.47), (3.41). The reduced problem (3.55)–(3.58) is uniquely solvable because U_2 is nonsingular (if the assumptions of Theorem 3.1 hold).

Using (3.47), (3.41) and Lemmas 2.1, 2.2 we get for the limits of $y(t,\varepsilon)$, $z(t,\varepsilon)$:

$$y(\infty,\varepsilon)=-A^{-1}(\infty,0)(B(\infty,0)\bar{z}(\infty)+f(\infty,0))+O(\varepsilon), \tag{3.61}$$

$$\begin{aligned}z(\infty,\varepsilon)=&-\big(D(\infty,0)-C(\infty,0)A^{-1}(\infty,0)B(\infty,0)\big)^{-1}\big(g(\infty,0)\\&-C(\infty,0)A^{-1}(\infty,0)f(\infty,0)\big)+O(\varepsilon).\end{aligned} \tag{3.62}$$

For the numerical solution of (3.1)–(3.4) we cut the infinite interval $[1,\infty]$ at a finite point $T\gg 1$ and replace the continuity requirement (3.4) by $r_++\tilde{r}_+$ boundary conditions at $t=T$. These asymptotic boundary conditions shall reflect the asymptotic behavior of y, z as $t\to\infty$. So we get the finite singular perturbation problem

$$\varepsilon y_T'=t^\alpha A(t,\varepsilon)y_T+t^\alpha B(t,\varepsilon)z_T+f(t,\varepsilon), \tag{3.63}$$

$$z_T'=t^\alpha C(t,\varepsilon)y_T+t^\alpha D(t,\varepsilon)z_T+g(t,\varepsilon), \qquad 1\le t\le T, \tag{3.64}$$

$$f(\varepsilon)\begin{pmatrix}y_T(1,\varepsilon)\\ z_T(1,\varepsilon)\end{pmatrix}=\beta(\varepsilon), \tag{3.65}$$

$$S(T,\varepsilon)\begin{pmatrix}y_T(T,\varepsilon)\\ z_T(T,\varepsilon)\end{pmatrix}=\gamma(T,\varepsilon) \tag{3.66}$$

where $S(T,\varepsilon)$ is an $(r_+ + \tilde{r}_+)\times(n+m)$-matrix, $\gamma(T,\varepsilon)\in\mathbb{R}^{r_+ + \tilde{r}_+}$ and y_T, x_T are the approximations to y and x. Asymptotic boundary conditions (for unperturbed problems) were constructed by de Hoog and Weiss (1980a), Markowich (1982b) and Lentini and Keller (1980).

Proceeding analogously, we set

$$S\equiv S(T,\varepsilon)=\begin{bmatrix} \left[I_{r_+},0\right]E^{-1}(\infty) & \left[E_{r_+},0\right]E^{-1}(\infty)A^{-1}(\infty,0)B(\infty,0) \\ 0 & \left[I_{\tilde{r}_+},0\right]E_2^{-1} \end{bmatrix}, \tag{3.67}$$

$$\gamma(T)\equiv\gamma(T,\varepsilon)=\begin{pmatrix} -\left[I_{r_+},0\right]E^{-1}(\infty)A^{-1}(\infty,0)f(T,0) \\ \left[I_{\tilde{r}_+},0\right]E_2^{-1}\bar{z}(\infty) \end{pmatrix}. \tag{3.68}$$

Let $\Omega(t,\varepsilon)$ denote the fundamental matrix of (the homogeneous system) (3.1), (3.2). Then, by proceeding as in de Hoog and Weiss (1980a), it is easy to show that $S\Omega(t,\varepsilon)$ does not contain exponentially increasing terms (as $t\to\infty$, $\varepsilon\to 0$). Therefore the boundary condition

$$S\begin{pmatrix} y(T,\varepsilon) \\ z(T,\varepsilon) \end{pmatrix}=0$$

sets the exponentially increasing solution components of the homogeneous problem (3.1), (3.2) to zero. $\gamma(T)$ is the (boundary) correction term for the inhomogeneous problem.

We now assume that the assumptions of Theorem 3.1 hold. As in the references cited above we get the following estimate for the unique solution y_T, x_T of (3.63)–(3.66):

$$\left\|\begin{pmatrix} y-y_T \\ z-z_T \end{pmatrix}\right\|_{[1,T]}\le K\left\|S\begin{pmatrix} y(T,\varepsilon) \\ z(T,\varepsilon) \end{pmatrix}-\gamma(T)\right\| \tag{3.69}$$

(for T sufficiently large and ε sufficiently small, but T and ε independent) where K is independent of T and ε. Using (3.59), (3.60) we get the convergence estimate

$$\left\|\begin{pmatrix} y-y_T \\ z-z_T \end{pmatrix}\right\|_{[1,T]}\le\text{const}(\|\bar{z}(T)-\bar{z}(\infty)\|+O(\varepsilon)) \tag{3.70}$$

where the constant is independent of T and ε.

Since the solutions of the reduced problem (3.63), (3.64) do not generally fulfill (3.66), y_T has a boundary large at $t=T$ whose height can be estimated by the right-hand side of (3.70). Estimates of the first term on the right-hand side of (3.70) depending on the decay of f, g as $t\to\infty$ are given in Markowich (1982b), (1983).

Under the assumptions of Theorem 3.1, the asymptotic boundary condition (3.66) can be constructed with respect to the perturbed infinite problem.

4. Quasilinear problems. We investigate

$$\varepsilon y'=t^\alpha A(z,t)y+t^\alpha f(z,y,t,\varepsilon), \tag{4.1}$$

$$z'=t^\alpha g(z,y,t,\varepsilon), \qquad \alpha>-1,\quad 1\le t<\infty, \tag{4.2}$$

$$f(\varepsilon)\begin{pmatrix} y(1,\varepsilon) \\ z(1,\varepsilon) \end{pmatrix}=\beta(\varepsilon), \tag{4.3}$$

$$\begin{pmatrix} y \\ z \end{pmatrix}\in C([1,\infty]) \tag{4.4}$$

where $A(z,t)$ is an $n\times n$-matrix, f an n-vector, g an m-vector and the problem (4.1), (4.2) is quasilinear:

$$\frac{\partial f}{\partial y}=O(\varepsilon) \tag{4.5}$$

for $t\in[1,\infty]$, $\varepsilon\in[0,\varepsilon_0]$ and y, z in compact sets. We get immediately

$$f(z,y,t,0)=f(z,0,t,0). \tag{4.6}$$

We now assume that $F(\varepsilon)$ is a $k\times(n+m)$-matrix (k will be specified later), $f,\beta\in C([0,\varepsilon_0])$ and Lipschitz continuous at $\varepsilon=0$ and

(4.7a) $f,g\in C^2(\mathbb{R}^{m+n}\times[1,\infty]\times[0,\varepsilon_0])$,

$f(z,y,\cdot,\varepsilon)\in C^1_\alpha([1,\infty])$ uniformly in compact subsets of $\mathbb{R}^{n+m}\times[0,\varepsilon_0]$,

(4.7b) $A\in C^2(\mathbb{R}^m\times[1,\infty])$, $A(z,\cdot)\in C^1_\alpha([1,\infty])$ uniformly in compact subsets of $\mathbb{R}^m$.

We define (the boundary conditions for) the reduced problems first since we will construct a solution to (4.1)–(4.4) which corresponds to a reduced solution. Therefore we proceed similarly to the linear case (see also Ringhofer (1981)). We split $F(0)$ as in (3.51)

$$F(0)=[\underbrace{F_y(0)}_{n},\underbrace{F_z(0)}_{m}], \tag{4.8}$$

and we assume that there is an integer $r_-\le n$ such that

$$F_y(0)=[\underbrace{F_{y_+}(0)}_{r_+},\underbrace{F_{y_-}(0)}_{r_-}], \qquad r_++r_-=n, \tag{4.9}$$

and the $k\times r_-$-matrix $F_{y_-}(0)$ ($k\ge r_-$ is assumed) has maximal rank r_-. Therefore, there is a $k\times k$-matrix Y such that

$$Yf_{y_-}(0)=\underbrace{\begin{bmatrix}V\\0\end{bmatrix}}_{r_-}\begin{matrix}\}r_-\\\}k-r_-\end{matrix}, \qquad Y=\underbrace{\begin{bmatrix}Y_1\\Y_2\end{bmatrix}}_{k}\begin{matrix}\}r_-\\\}k-r_-\end{matrix} \tag{4.10}$$

holds where V is nonsingular.

The main assumption is that the reduced problem

$$\bar z'=t^\alpha g(\bar z,\bar y,t,0), \qquad 1\le t<\infty, \tag{4.11}$$

$$0=A(\bar z,t)y+f(\bar z,0,t,0), \qquad 1\le t<\infty, \tag{4.12}$$

$$Y_2F(0)\begin{pmatrix}\bar y(1)\\ \bar z(1)\end{pmatrix}=Y_2\beta(0), \tag{4.13}$$

$$\begin{pmatrix}\bar y\\ \bar z\end{pmatrix}\in C([1,\infty]) \tag{4.14}$$

has an isolated solution (see Keller (1975)), $\bar y$, $\bar z$, and that

$$A(z,t)=\left[\begin{array}{c|c}\underbrace{A_+(z,t)}_{r_+} & \underbrace{0}_{r_-}\\ \hline \underbrace{0}_{r_+} & \underbrace{A_-(z,t)}_{r_-}\end{array}\right]\begin{matrix}\}r_+\\\}r_-\end{matrix} \tag{4.15}$$

holds in $C_\varphi=\{(z,t)|\,\|z-\bar z(t)\|\le\varphi,\ t\in[1,\infty]\}$, $\varphi>0$ sufficiently large, where the eigenvalues $\lambda_+(z,t)$ of $A_+(z,t)$ and $\lambda_-(z,t)$ of $A_-(z,t)$ fulfill

$$\mathrm{Re}\lambda_+(z,t)\ge c_+>0, \qquad (z,t)\in C_\varphi, \tag{4.16}$$

$$\mathrm{Re}\lambda_-(z,t)\le -c_-<0, \qquad (z,t)\in C_\varphi. \tag{4.17}$$

This guarantees that (4.12) can be solved with respect to y for $z\in C_\varphi$ and

$$\bar y=\bar y(\bar z,t)=-A^{-1}(\bar z,t)f(\bar z,0,t,0) \tag{4.18}$$

holds. We assume that the matrix

$$\tilde D=\frac{\partial g}{\partial z}(\bar z(\infty),\bar y(\infty),\infty,0)+\frac{\partial g}{\partial y}(\bar z(\infty),\bar y(\infty),\infty,0)\frac{\partial \bar y}{\partial \bar z}(\bar z(\infty),\infty) \tag{4.19}$$

fulfills

$$\tilde D=E\tilde J E^{-1}, \qquad \tilde J=\begin{bmatrix}\tilde J^+ & 0\\ 0 & \tilde J^-\end{bmatrix}\begin{matrix}\}\tilde r_+\\ \}\tilde r_-\end{matrix} \tag{4.20}$$

where the eigenvalues of $\tilde J^+$ ($\tilde J^-$) have positive (negative) real parts.

Therefore we assume that $Y_2 f(0)$ is an $\tilde r_-\times(n+m)$-matrix, $Y_2\beta(0)\in\mathbb{R}^{r_-}$ and $k=r_-+\tilde r_-$, such that (4.11), (4.12), (4.13), (4.14) is well posed with respect to the number of "finite" boundary conditions (see Markowich (1983) and de Hoog and Weiss (1980a, b)). Obviously $z_\infty=\bar z(\infty)$, $y_\infty=\bar y(\infty)$ are solutions of

$$0=g(z_\infty,y_\infty,\infty,0), \tag{4.21a}$$

$$0=A(z_\infty,\infty)y_\infty+f(z_\infty,0,\infty,0), \tag{4.21b}$$

and we assume that z_∞, y_∞ are isolated. Therefore $\tilde D$ as of (4.19) can be calculated a priori at these roots.

Let $\psi(t,\varepsilon)$ denote the fundamental matrix of

$$\varepsilon v'=t^\alpha A(\bar z(t),t)v, \qquad \psi(1,\varepsilon)=I. \tag{4.22}$$

We only state the existence result since the proof goes along the lines of the proof given in Ringhofer (1981) for finite interval problems using the linear theory developed in §3 of this paper.

THEOREM 4.1. *Let $f(\varepsilon)$ be an $(r_-+\tilde r_-)\times(n+m)$-matrix. Assume that f, β are Lipschitz continuous at $\varepsilon=0$, that* (4.7), (4.9) *hold and that the reduced problem* (4.11)–(4.14) *has an isolated solution $\bar y$, $\bar z$ such that* (4.16), (4.17) *hold and $f(z_\infty,y_\infty,t,0)\equiv 0$, $g(z_\infty,y_\infty,t,0)\equiv 0$ for $t\ge\delta\ge 1$ where z_∞, y_∞ fulfill* (4.21). *Then the problem* (4.1), (4.2), (4.3), (4.4) *has a locally unique solution y, z for ε sufficiently small such that*

$$y(t,\varepsilon)=\psi(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\zeta+\bar y(t)+O(\varepsilon),$$

$$z(t,\varepsilon)=\bar z(t)+O(\varepsilon)$$

holds uniformly in $[1,\infty]$ for some $\zeta\in\mathbb{C}^{r_-}$.

From §3 we conclude that

$$\left\|\psi(t,\varepsilon)\begin{bmatrix}0\\ I_{r_-}\end{bmatrix}\right\|\le \text{const}\cdot\exp\left(\frac{-c}{\varepsilon(\alpha+1)}(t^{\alpha+1}-1)\right), \qquad c>0, \tag{4.23}$$

holds. t-asymptotics for $\bar{z}(t), \bar{y}(t)$ can be obtained from Markowich (1983):

$$\bar{z}(t)=\bar{z}(\infty)+E\phi(t,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi+O\left(\left\|\phi(t,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\right\|^2\right) \tag{4.24}$$

for $\xi\in\mathbb{C}^{\tilde{r}_-}$ where

$$\phi(t,1)=\exp\left(\frac{\tilde{J}}{\alpha+1}(t^{\alpha+1}-1)\right) \tag{4.25}$$

holds. From (4.19) we get

$$\bar{y}(t)=\bar{y}(\infty)+\frac{\partial\bar{y}}{\partial\bar{z}}(\bar{z}(\infty),\infty)E\phi(t,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\xi+O\left(\left\|\phi(t,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\right\|^2\right). \tag{4.26}$$

The approximating "finite" problems are

$$\varepsilon y_T'=t^\alpha A(z_T,t)y_T+t^\alpha f(z_T,y_T,t,\varepsilon), \tag{4.27}$$

$$z_T'=t^\alpha g(y_T,z_T,t,\varepsilon), \qquad 1\le t\le T, \tag{4.28}$$

$$f(\varepsilon)\begin{pmatrix}y_T(1,\varepsilon)\\ z_T(1,\varepsilon)\end{pmatrix}=\beta(\varepsilon), \tag{4.29}$$

$$S(T,\varepsilon)\begin{pmatrix}y_T(T,\varepsilon)\\ z_T(T,\varepsilon)\end{pmatrix}=\gamma(T,\varepsilon) \tag{4.30}$$

where $S(T,\varepsilon)$ is an $(r_+ +\tilde{r}_+)\times(n+m)$-matrix and $\gamma(T,\varepsilon)\in\mathbb{R}^{r_+ +\tilde{r}_+}$. We choose

$$S\equiv S(T,\varepsilon)=\begin{bmatrix}[I_{r_+},0] & -[I_{r_+},0]\dfrac{\partial\bar{y}}{\partial\bar{z}}(\bar{z}(\infty),\infty)\\ 0 & [I_{\tilde{r}_+},0]E^{-1}\end{bmatrix} \tag{4.31}$$

and

$$\gamma\equiv\gamma(T,\varepsilon)=S\begin{pmatrix}\bar{y}(\infty)\\ \bar{z}(\infty)\end{pmatrix}. \tag{4.32}$$

Then we obtain

$$\left\|S\begin{pmatrix}y(T,\varepsilon)\\ z(T,\varepsilon)\end{pmatrix}-\gamma\right\|=O\left(\left\|\phi(T,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\right\|^2\right)+O(\varepsilon). \tag{4.33}$$

Using the linear stability result (3.69), we get by proceeding as de Hoog and Weiss (1980a) did:

$$\left\|\begin{pmatrix}y_T-y\\ z_T-z\end{pmatrix}\right\|_{[1,T]}\le\text{const}\left(\left\|\phi(T,1)\begin{bmatrix}0\\ I_{\tilde{r}_-}\end{bmatrix}\right\|^2+O(\varepsilon)\right) \tag{4.34}$$

for the locally unique solution y_T, z_T of (4.27), (4.28), (4.29), (4.30) such that (4.33) constitutes the convergence estimate.

As in the linear case, this asymptotic boundary condition only depends on the reduced "infinite" problem.

REFERENCES

[1] U. ASCHER AND R. WEISS (1981), *Collocation for singular perturbation problems* I, *First order systems with constant coefficients*, TR 81-2, Univ. British Columbia, Vancouver, BC, Canada.

[2] F. DE HOOG AND R. WEISS (1980a), *An approximation method for boundary value problems on infinite intervals*, Computing, 24, pp. 227–239.

[3] _____ (1980b), *On the boundary value problem for systems of ordinary differential equations with a singularity of the second kind*, this Journal, 11, pp. 41–60.

[4] F. C. HOPPENSTAEDT (1966), *Singular perturbation problems on the infinite interval*, Trans. Amer. Math. Soc., 123, pp. 521–535.

[5] H. O. KREISS AND N. NICHOLS (1975), *Numerical methods for singular perturbation problems*, Report 57, Dept. of Computer Science, Uppsala University, Uppsala, Sweden.

[6] P. A. LAGERSTROM (1961), *Méthodes asymptotiques pour l'étude des equations de Navier–Stokes*, Lecture Notes, Institut Henri Poincaré, Paris.

[7] P. A. LAGERSTROM AND R. G. CASTEN (1972), *Basic concepts underlying singular perturbation techniques*, SIAM Rev., 14, pp. 63–120.

[8] M. LENTINI AND H. B. KELLER (1980), *Boundary value problems on semi-infinite intervals and their numerical solution*, SIAM J. Numer. Anal., 17, pp. 577–604.

[9] R. E. O'MALLEY, JR. (1978), *On singular singularly perturbed initial value problems*, Applicable Anal., 8, pp. 71–81.

[10] _____ (1979), *A singular singularly perturbed linear boundary value problem*, this Journal, 10, pp. 695–709.

[11] P. A. MARKOWICH (1982a), *Asymptotic analysis of von Karman flows*, SIAM J. Appl. Math., 42, pp. 549–557.

[12] _____ (1982b), *A theory for the approximation of solutions of boundary value problems on infinite intervals*, this Journal, 13 (1982), pp. 484–513.

[13] _____ (1983), *Analysis of boundary value problems on infinite intervals*, this Journal, 14, pp. 11–37.

[14] P. A. MARKOWICH AND C. A. RINGHOFER (1981), *The numerical solution of boundary value problems on long intervals*, MRC Tech. Sum. Rep. 2205, Mathematics Research Center, Univ. of Wisconsin, Madison.

[15] C. A. RINGHOFER (1980), *Collocation methods for singularly perturbed boundary value problems*, Master's thesis, Technische Universität Wien, Vienna, Austria.

[16] _____ (1981), *A class of collocation schemes for singularly perturbed boundary value problems*, Thesis, Technische Universität Wien, Vienna, Austria.

[17] H. SCHLICHTING (1959), *Entstehung der Turbulenz*, in Fluid Dynamics 1, Handbuch der Physik, S. Flügge, ed., Springer-Verlag, Berlin.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0006 $01.25/0

EQUIPARTITION OF ENERGY IN SCATTERING THEORY*

GEORGE DASSIOS† AND MANOUSOS GRILLAKIS‡

Abstract. It is shown that the difference between the kinetic and the potential energy of a solution of the wave equation in the exterior of a star-shaped body, which vanishes on the surface of the body and has Cauchy data with compact support, decays to zero as time tends to infinity. Therefore asymptotic equipartition of energy occurs even in the presence of a scatterer. An upper bound for the rate of decay has been found. An extension of Morawetz's local energy decay result for the case of an expanding sphere is also obtained.

1. Introduction. Equipartition of energy has been studied for many physically interesting [2]–[6] as well as abstract [1], [8], [9], [10] cases in the form of initial value problems for hyperbolic equations in $\mathbb{R}^3$. The basic technique to prove equipartion (or partition) of energy, in most cases, is to Fourier (or Radon) transform the problem and use Paley–Weiner type theorems and Parseval's identity.

Nevertheless, if a scatterer is present the fundamental domain of the solution does not contain all of $\mathbb{R}^3$. A region $\mathcal{B}$ (the scatterer) is left out of $\mathbb{R}^3$ and boundary conditions on $\partial\mathcal{B}$ must be prescribed. As a consequence, the transform technique is not applicable anymore, and a new method should be used to investigate equipartion of energy for this problem.

In this paper we prove an asymptotic equipartition of energy result for the case of the classical wave equation when the scatterer is star-shaped, the wave vanishes on the surface of the scatterer and the initial data have compact support. The asymptotic character of the result is a consequence of the Huygens' principle which in the presence of a body ceases to hold. The main idea of the paper is to look at the decay properties of the solution and its derivatives, locally, as the wave propagates along the characteristics. In order to obtain the necessary pointwise estimates we extend the analysis given by Friedlander in his fundamental papers [7, Parts I, II]. This local study of the solution along the characteristics does not take into consideration the influence from past reflections on the boundary. Therefore the scattering process is isolated from the past for every outgoing wave. There was also need to generalize Morawetz's local energy decay result to a sphere which extends at a speed strictly less than the phase velocity of the wave. It was found that even in this case, where the domain of integration of the energy expands, the rate of decay is at least as fast as t^{-1}.

2. Pointwise estimates. Consider a closed, simply connected and star-shaped subset $\mathcal{B}$ of $\mathbb{R}^3$, bounded by the smooth surface $\partial\mathcal{B}$. Let $u(\mathbf{x},t)$ be a C^2-solution of the wave equation

$$\Box u(\mathbf{x},t)=u_{tt}(\mathbf{x},t)-\Delta u(\mathbf{x},t)=0, \qquad \mathbf{x}\in\mathbb{R}^3-\mathcal{B}=\mathcal{B}^c, \quad t\geq 0 \tag{1}$$

which satisfies the boundary condition

$$u(\mathbf{x},t)=0, \qquad \mathbf{x}\in\partial\mathcal{B}, \quad t\geq 0, \tag{2}$$

and the initial conditions

$$u(\mathbf{x},0)=f(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{B}^c, \tag{3}$$

$$u_t(\mathbf{x},0)=g(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{B}^c. \tag{4}$$

* Received by the editors December 18, 1981.
† Department of Mathematics, University of Patras, Greece.
‡ National Technical University of Athens, Greece.

Assume that the initial data f and g are smooth functions with compact supports and that

$$(\text{supp} f)\cup(\text{supp} g)\subset B(\mathbf{0},a) \tag{5}$$

where $B(\mathbf{0},a)$ is the open ball centered at $\mathbf{0}$ with radius a. Since the support of a function is a closed set, (5) implies that both f and g vanish on the surface of the sphere $S(\mathbf{0},a)=\partial B(\mathbf{0},a)$.

Following Friedlander [7], we translate the time axis by $-a$ so that the solution u vanishes on the characteristic surface (cone)

$$t=|\mathbf{x}|=r. \tag{6}$$

The initial-boundary value problem (1)–(4) then reads as follows: Find a C^2 solution of

$$\Box u(\mathbf{x},t)=0, \qquad (\mathbf{x},t)\in\Omega, \tag{7}$$

which satisfies the conditions

$$u(\mathbf{x},t)=0, \qquad \mathbf{x}\in\partial\mathcal{B}\cup\{\mathbf{x}\in\mathbb{R}^3: t=r=|\mathbf{x}|\}, \quad t\geq a, \tag{8}$$

$$u(\mathbf{x},a)=f(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{B}^c\cap B(\mathbf{0},a), \tag{9}$$

$$u_t(\mathbf{x},a)=g(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{B}^c\cap B(\mathbf{0},a), \tag{10}$$

where Ω is the shaded region of Fig. 1. Since the solution is zero in the exterior of $B(\mathbf{0},a)$ at $t=a$, the following Kirchhoff's integral representation holds for $(\mathbf{x},t)$ in the shaded region of Fig. 2 determined by the inequalities $t\geq r\geq a$,

$$u(\mathbf{x},t)=\frac{a^2}{4\pi}\int_\omega\left\{-\frac{1}{R}u_r(a\hat{\boldsymbol{\xi}},t-R)+\frac{\mathbf{x}\cdot\hat{\boldsymbol{\xi}}-a}{R^2}\left[\frac{1}{R}u(a\hat{\boldsymbol{\xi}},t-R)+u_t(a\hat{\boldsymbol{\xi}},t-R)\right]\right\}d\omega(\hat{\boldsymbol{\xi}}) \tag{11}$$

where ω stands for the surface of the unit sphere in $\mathbb{R}^3$, $\hat{\boldsymbol{\xi}}$ is the exterior unit normal on ω and

$$R=|\mathbf{x}-a\hat{\boldsymbol{\xi}}|. \tag{12}$$

LEMMA 1. *If u is a C^2-solution of* (7)–(10) *then*

$$u_t(\mathbf{x},t)=O\left(\frac{1}{r}\right), \qquad r\to\infty, \tag{13}$$

whenever $t-r=\tau$, is positive and bounded.

Proof. The smoothness of the integrand in Kirchhoff's formula (11) allows a differentiation under the integral sign. Hence the time derivative of (11) gives

$$u_t(\mathbf{x},t)=\frac{a^2}{4\pi}\int_\omega\left\{-\frac{1}{R}u_{rt}(a\hat{\boldsymbol{\xi}},t-R)+\frac{\mathbf{x}\cdot\hat{\boldsymbol{\xi}}-a}{R^2}\left[\frac{1}{R}u_t(a\hat{\boldsymbol{\xi}},t-R)+u_{tt}(a\hat{\boldsymbol{\xi}},t-R)\right]\right\}d\omega(\hat{\boldsymbol{\xi}}). \tag{14}$$

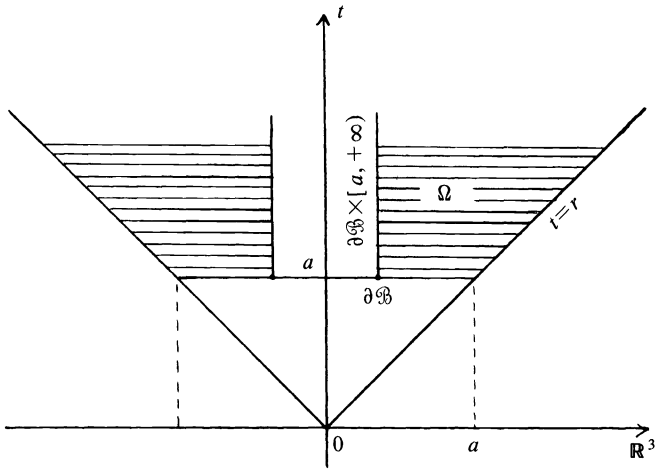


FIG. 1

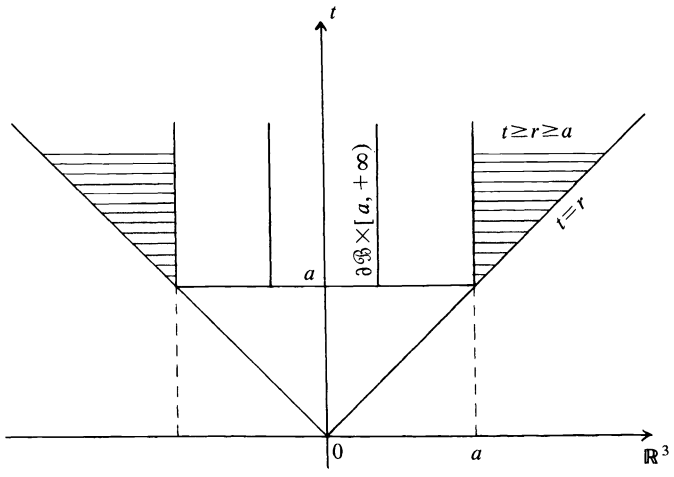


FIG. 2. $R=|\mathbf{x}-a\hat{\xi}|$.

Using the basic order relations for $r\to\infty$:

$$\frac{r}{R}=1+O\left(\frac{1}{r}\right), \tag{15}$$

$$R=r-a\cos\gamma+O\left(\frac{1}{r}\right), \tag{16}$$

$$\frac{r\cos\gamma-a}{R}=\cos\gamma+O\left(\frac{1}{r}\right), \tag{17}$$

where

$$\mathbf{x}\cdot\hat{\xi}=r\cos\gamma, \tag{18}$$

and the fact that $t-R$ is bounded because of (16), which implies that the integrand in (14) is a continuous function over a compact set, we obtain from (14) the following estimate:

(19)

$$u_t(\mathbf{x},t)=\frac{a^2}{4\pi r}\int_\omega\Big\{-u_{rt}(a\hat{\xi},t-R)\Big(1+O\Big(\frac{1}{r}\Big)\Big)+\Big(1+O\Big(\frac{1}{r}\Big)\Big)\cdot\Big(\cos\gamma+O\Big(\frac{1}{r}\Big)\Big)\Big[\Big(1+O\Big(\frac{1}{r}\Big)\Big)\frac{1}{r}u_t(a\hat{\xi},t-R)+u_{tt}(a\hat{\xi},t-R)\Big]\Big\}d\omega(\hat{\xi})$$

$$=\frac{a^2}{4\pi r}\int_\omega\left[-u_{rt}(a\hat{\xi},t-R)+\cos\gamma u_{tt}(a\hat{\xi},t-R)\right]d\omega(\hat{\xi})+O\Big(\frac{1}{r^2}\Big)=O\Big(\frac{1}{r}\Big).$$

This proves Lemma 1.

LEMMA 2. *If the assumptions of Lemma* 1 *hold then*

$$|\nabla u(\mathbf{x},t)|=O\Big(\frac{1}{r}\Big),\qquad r\to\infty. \tag{20}$$

Proof. Let I be the integrand of the representation (11). Then by straightforward calculations we obtain

$$\frac{\partial I}{\partial x_i}=I_i=\frac{1}{R}\cdot\frac{x_i-a\xi_i}{R}\Big[\frac{1}{R}u_r+u_{rt}\Big]+\Big[\frac{1}{R^2}\cdot\frac{\xi_i}{R}-\frac{3}{R^3}\cdot\frac{\mathbf{x}\cdot\hat{\xi}-a}{R}\cdot\frac{x_i-a\xi_i}{R}\Big]\Big[\frac{1}{R}u+u_t\Big]-\frac{1}{R}\cdot\frac{\mathbf{x}\cdot\hat{\xi}-a}{R}\cdot\frac{x_i-a\xi_i}{R}u_{tt} \tag{21}$$

where the function u and its derivatives are evaluated at the point $(a\hat{\xi},t-R)$ and $i=1,2,3$. Using the relation

$$\frac{x_i-a\xi_i}{R}=\frac{x_i}{r}+O\Big(\frac{1}{r}\Big),\qquad r\to\infty, \tag{22}$$

in (21) and the continuity of the second derivatives of u we obtain

$$\lim_{r\to\infty}rI_i=\frac{x_i}{r}\left[u_{rt}(a\hat{\xi},\tau+a\cos\gamma)-\cos\gamma u_{tt}(a\hat{\xi},\tau+a\cos\gamma)\right] \tag{23}$$

where

$$\tau+a\cos\gamma=\lim_{r\to\infty}(t-R). \tag{24}$$

From (11) we obtain

$$\lim_{r\to\infty}r^2|\nabla u|^2=\frac{a^2}{4\pi}\sum_{i=1}^{3}\lim_{r\to\infty}\left[\int_\omega rI_i\,d\omega\right]^2=\frac{a^2}{4\pi}\sum_{i=1}^{3}\left[\int_\omega(\lim rI_i)\,d\omega\right]^2 \tag{25}$$

where the passage of the limit inside the integral is justified by using the Lebesgue dominated convergence theorem. Relation (23) implies that the last side of equation (25) is bounded. Therefore

$$|\nabla u|^2 = O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{26}$$

from which (20) is obtained, and the proof of Lemma 2 is completed.

LEMMA 3. *Under the assumptions of Lemma* 1 *it is true that*

$$|u_t(\mathbf{x},t)\hat{\mathbf{x}} + \nabla u(\mathbf{x},t)| = O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{27}$$

where $\hat{\mathbf{x}} = r^{-1}\mathbf{x}$.

Proof. We use (14) and (21) to evaluate the ith component of the vector field $u_t\hat{\mathbf{x}} + \nabla u$

$$u_t \frac{x_i}{r} + u_{x_i} = \frac{a^2}{4\pi}\int_\omega \zeta_i \, d\omega \tag{28}$$

where

$$\begin{aligned}
\zeta_i = {} & \frac{1}{R}\left[\frac{x_i - a\xi_i}{R} - \frac{x_i}{r}\right] u_{rt}(a\hat{\xi}, t-R) \\
& - \frac{1}{R}\cdot\frac{\mathbf{x}\cdot\hat{\xi} - a}{R}\left[\frac{x_i - a\xi_i}{R} - \frac{x_i}{r}\right] u_{tt}(a\hat{\xi}, t-R) \\
& + \frac{1}{R^2}\left[\frac{\mathbf{x}\cdot\hat{\xi} - a}{R}\left(\frac{x_i}{r} - 3\,\frac{x_i - a\xi_i}{R}\right) + \xi_i\right] u_t(a\hat{\xi}, t-R) \\
& + \frac{1}{R^2}\cdot\frac{x_i - a\xi_i}{R} u_r(a\hat{\xi}, t-R) \\
& + \frac{1}{R^3}\left[\xi_i - 3\,\frac{\mathbf{x}\cdot\hat{\xi} - a}{R}\cdot\frac{x_i - a\xi_i}{R}\right] u(a\hat{\xi}, t-R).
\end{aligned} \tag{29}$$

If we use the relations (15)–(18) and (22) to estimate the coefficients of u_{rt}, u_{tt}, u_t and u_r we observe that, with the exception of the coefficient of u which is $O(1/r^3)$, all other coefficients are $O(1/r^2)$. Therefore

$$\zeta_i = O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{30}$$

for every $i = 1, 2, 3$. If we combine now the continuity of the second derivatives of u, the boundedness of $t - r$ and the estimate (30) we obtain from (28) and (29) that

$$u_t \frac{x_i}{r} + u_{x_i} = O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{31}$$

for each $i = 1, 2, 3$. Equation (27) is then immediate and the proof of Lemma 3 is completed.

3. Energy estimates. We next turn to global estimates for the solution u. Morawetz [11] has proved that if u is a solution of (1)–(5) and $\mathfrak{B}$ is star-shaped, then the energy that is contained in any sphere of constant radius decays at least as fast as t^{-1}, as the time t tends to infinity.

In this paper we extend Morawetz's result and prove that the same rate of decay holds even when the sphere, where the energy is evaluated, is expanding with a speed not exceeding the phase velocity of the wave. More precisely, we will prove the following theorem.

THEOREM 1. *If u is a solution of* (5), (7)–(10), *and $\mathcal{B}$ is star-shaped, then*

$$\int_{V_{\mathcal{B}}^{\lambda t+a(1-\lambda)}} \left(|u_t|^2+|\nabla u|^2\right) d^3x = O\left(\frac{1}{t}\right), \qquad t\to\infty, \tag{32}$$

where $V_{\mathcal{B}}^{\lambda t+a(1-\lambda)}$ is the region exterior to the scatterer $\mathcal{B}$ and interior to the ball $B(\mathbf{0},\lambda t+a(1-\lambda))$ and $\lambda\in[0,1)$.

Proof. Consider the following identity:

$$\frac{\partial}{\partial t}\left\{\frac{t}{2}\left(u_t^2+|\nabla u|^2\right)+ru_ru_t+uu_t\right\} + \nabla\cdot\left\{-\frac{\mathbf{x}}{2}\left(u_t^2-|\nabla u|^2\right)-(u+tu_t+ru_r)\nabla u\right\}=0. \tag{33}$$

We next integrate (33) over the region

$$\Omega_{\mathcal{B}}^k=[\mathcal{B}^c\cap B(\mathbf{0},k)]\times[a,t]\subset\mathbb{R}^3\times[0,+\infty) \tag{34}$$

where the radius k of the ball $B(\mathbf{0},k)$ is chosen so that $u(\mathbf{x},t')=0$ for every $r\geq k$ and $a\leq t'\leq t$. $\Omega_{\mathcal{B}}^k$ is the shaded region of Fig. 3. By Gauss' theorem and the particular choice of k, the above integration gives

$$\begin{aligned}\int_a^t\int_{\partial\mathcal{B}}\left[-\frac{\mathbf{x}}{2}\left(u_{t'}^2-|\nabla u|^2\right)-(u+t'u_{t'}+ru_r)\nabla u\right]\cdot\hat{\mathbf{n}}_x\,ds\,dt \\ +\int_{\mathcal{B}^c}\left[\frac{t'}{2}\left(u_{t'}^2+|\nabla u|^2\right)+ru_ru_{t'}+uu_{t'}\right]_{t'=t}d^3x \\ -\int_{\mathcal{B}^c}\left[\frac{t'}{2}\left(u_{t'}^2+|\nabla u|^2\right)+ru_ru_{t'}+uu_{t'}\right]_{t'=a}d^3x=0,\end{aligned} \tag{35}$$

where $\hat{\mathbf{n}}_x$ is the interior unit normal on $\partial\mathcal{B}$.

Since $u=0$ on $\partial\mathcal{B}$, the gradient of u has the direction of the normal on the surface of the scatterer. Hence for $\mathbf{x}\in\partial\mathcal{B}$ we have

$$\left[-\frac{\mathbf{x}}{2}\left(u_{t'}^2-|\nabla u|^2\right)-(u+t'u_{t'}+ru_r)\nabla u\right]\cdot\hat{\mathbf{n}}_x=-\frac{\mathbf{x}\cdot\hat{\mathbf{n}}_x}{2}|\nabla u|^2\geq 0 \tag{36}$$

where the last inequality is justified by the star-shaped property of the body which ensures that

$$\mathbf{x}\cdot\hat{\mathbf{n}}_x\leq 0, \qquad \mathbf{x}\in\partial\mathcal{B}. \tag{37}$$

Therefore the first integral in (35) is nonnegative. Since the initial data are smooth and have compact support, the last integral in (35) is bounded. Relation (35) then gives for the time t the following inequality:

$$\int_{\mathcal{B}^c}\left[\frac{t}{2}\left(u_t^2+|\nabla u|^2\right)+ru_ru_t+uu_t\right]d^3x\leq M_1 \tag{38}$$

where M_1 is a bound of the last integral in (35). Introduce the notation V_a^β for the space between the balls $B(\mathbf{0},a)$ and $B(\mathbf{0},\beta)$ with $a<\beta$, and $V_{\mathcal{B}}^\beta$ for the space between the

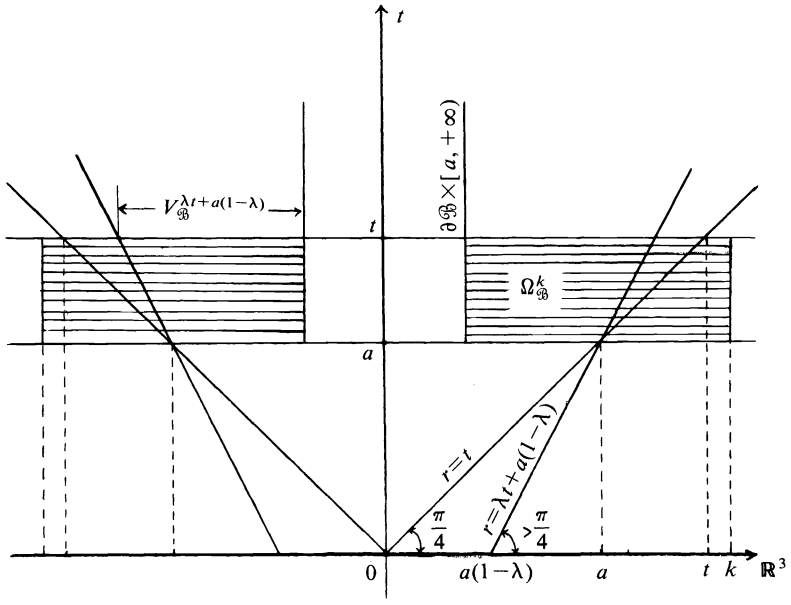


FIG. 3

surface $\partial\mathfrak{B}$ of the scatterer and the ball $B(\mathbf{0},\beta)$ which includes the scatterer. Inequality (38) can be written as follows:

$$K_1+K_2+K_3+K_4+K_5\leq M_1 \tag{39}$$

where

$$K_1=\int_{V_{\mathfrak{B}}^{\lambda t+a(1-\lambda)}} ru_r u_t d^3x, \tag{40}$$

$$K_2=\int_{V_{\mathfrak{B}}^{\lambda t+a(1-\lambda)}} \frac{t}{2}\left(u_t^2+|\nabla u|^2\right) d^3x, \tag{41}$$

$$K_3=\int_{V_{\lambda t+a(1-\lambda)}^{t}} ru_r u_t d^3x, \tag{42}$$

$$K_4=\int_{V_{\lambda t+a(1-\lambda)}^{t}} \frac{t}{2}\left(u_t^2+|\nabla u|^2\right) d^3x, \tag{43}$$

$$K_5=\int_{\mathfrak{B}^c} uu_t d^3x. \tag{44}$$

For K_5 we have the following bound (Morawetz [11, Lemma 3c]):

$$|K_5|\leq M_2. \tag{45}$$

For K_1 we obtain

$$|K_1|\leq\frac{\lambda t+a(1-\lambda)}{2}\int_{V_{\mathfrak{B}}^{\lambda t+a(1-\lambda)}}\left(u_t^2+|\nabla u|^2\right) d^3x \tag{46}$$

$$\leq\frac{\lambda t}{2}\int_{V_{\beta}^{\lambda t+a(1-\lambda)}}\left(u_t^2+|\nabla u|^2\right) d^3x+M_3.$$

Also for K_3 we have that

$$|K_3| \leq \frac{t}{2} \int_{V^t_{\lambda t + a(1-\lambda)}} \left(u_t^2 + |\nabla u|^2 \right) d^3x = K_4. \tag{47}$$

Combine the inequalities (39)–(47) to obtain

$$\begin{aligned} K_2 &\leq M_1 - K_1 - K_3 - K_4 - K_5 \\ &\leq M_1 + |K_1| + |K_3| - K_4 + |K_5| \\ &\leq \lambda K_2 + M_1 + M_2 + M_3 \end{aligned} \tag{48}$$

or

$$\int_{V^{\lambda t + a(1-\lambda)}_{\mathfrak{B}}} \left(u_t^2 + |\nabla u|^2 \right) d^3x \leq \frac{M_1 + M_2 + M_3}{2(1-\lambda)} \cdot \frac{1}{t}, \tag{49}$$

which implies (32) and proves the theorem.

Remark. Theorem 1 reduces, for $\lambda = 0$, to Morawetz's [11] result for the local energy decay in a sphere of constant radius. The estimate (49) indicates that as λ increases, the constant $(M_1 + M_2 + M_3)/2(1-\lambda)$ becomes larger. In other words, the faster the sphere expands the larger the constant in (49) is. This is physically reasonable for all speeds of expansion $\lambda < 1$.

Our next theorem implies that the energy that crosses the section $t_1 \leq t \leq t_2$ of the characteristic cone $r = \lambda t + a(1-\lambda)$ decays at the rate $1/(t+A)$ where $-A$ is the point where the cone meets the time axis.

THEOREM 2. *Under the hypotheses of Theorem* 1 *it is true that*

$$\int_{V^{t_2}_{\mathfrak{B}} + A} \left[u_t^2 + |\nabla u|^2 \right]_{t=t_2} d^3x = \int_{V^{t_1}_{\mathfrak{B}} + A} \left[u_t^2 + |\nabla u|^2 \right]_{t=t_1} d^3x + O\left(\frac{1}{t_1 + A} \right), \qquad t_1 \to \infty \tag{50}$$

where $t_2 > t_1 > a$, $0 \leq \lambda < 1$ *and* $A = (\lambda - 1)t_1 + a(1-\lambda) < 0$.

Proof. Integrating the identity

$$(u_{tt} - \Delta u) u_t = \frac{\partial}{\partial t} \left[u_t^2 + |\nabla u|^2 \right] + \nabla \cdot [-2u_t \nabla u] = 0 \tag{51}$$

over the space-time region bounded by the planes $t = t_1 \geq a$, $t = t_2 > t_1$, the conical surface $r = t + A$ and the cylinder $\partial\mathfrak{B} \times [t_1, t_2]$, and applying Gauss' theorem we obtain

$$\begin{aligned} &\int_{V^{t_2}_{\mathfrak{B}} + A} \left[u_t^2 + |\nabla u|^2 \right]_{t=t_2} d^3x - \int_{V^{t_1}_{\mathfrak{B}} + A} \left[u_t^2 + |\nabla u|^2 \right] d^3x \\ &\quad = \frac{1}{\sqrt{2}} \int_{t_1}^{t_2} \int_{r=t+A} \left[u_t^2 + |\nabla u|^2 + 2u_t \hat{\mathbf{x}} \cdot \nabla u \right] ds\, dt \\ &\quad = \frac{1}{\sqrt{2}} \int_{t_1}^{t_2} \int_{r=t+A} |u_t \hat{\mathbf{x}} + \nabla u|^2 ds\, dt. \end{aligned} \tag{52}$$

By Lemma 3 we have that

$$(53)\qquad \int_{t_1}^{t_2}\int_{r=t+A}|u_t\hat{\mathbf{x}}+\nabla u|^2\,ds\,dt$$

$$=\int_{t_1}^{t_2}\int_{r=t+A}O\left(\frac{1}{r^4}\right)ds\,dt=O\left(\int_{t_1}^{t_2}\int_{r=t+A}\frac{1}{r^4}\,ds\,dt\right)$$

$$=O\left(\int_{t_1}^{t_2}\frac{1}{(t+A)^2}\,dt\right)=O\left(\frac{1}{t_1+A}-\frac{1}{t_2+A}\right)$$

$$=O\left(\frac{1}{t_1+A}\right)$$

where we have used the fact that $t_1<t_2$ and the property

$$(54)\qquad \int O(f)\,dx=O\left(\int|f|\,dx\right).$$

Relations (52) and (53) imply (50). The proof of Theorem 2 is completed.

LEMMA 4. *If u is a solution of* (5), (7)–(10), *and* $0<t+A_1<t+A_2$, *then*

$$(55)\qquad \int_{V_{t+A_1}^{t+A_2}}\left(u_t^2-|\nabla u|^2\right)d^3x=O\left(\ln\frac{t+A_2}{t+A_1}\right).$$

Proof. From Lemmas 1 and 2 we have that

$$(56)\qquad |u_t\hat{\mathbf{x}}-\nabla u|=O\left(\frac{1}{r}\right),$$

and by Lemma 3 and (56) we obtain

$$(57)\qquad u_t^2-|\nabla u|^2=(u_t\hat{\mathbf{x}}+\nabla u)\cdot(u_t\hat{\mathbf{x}}-\nabla u)\le|u_t\hat{\mathbf{x}}+\nabla u||u_t\hat{\mathbf{x}}-\nabla u|=O\left(\frac{1}{r^3}\right).$$

Therefore

$$(58)\qquad \int_{V_{t+A_1}^{t+A_2}}\left(u_t^2-|\nabla u|^2\right)d^3x=O\left(\int_{t+A_1}^{t+A_2}\int_{|\mathbf{x}|=r}\frac{1}{r^3}\,ds\,dr\right)=O\left(\ln\frac{t+A_2}{t+A_1}\right).$$

4. Equipartion of energy. In this section we state and prove our main result, which implies equipartition of energy when a scatterer is present.

THEOREM 3. *If u is a solution of* (5), (7)–(10), *then*

$$(59)\qquad \int_{\mathcal{B}^c}\left[u_t^2-|\nabla u|^2\right]d^3x=O\left(\frac{1}{t^{1/2}}\right)$$

as $t\to\infty$, i.e., asymptotic equipartition of energy is attained at the rate $t^{-1/2}$.

Proof. If we set

$$(60)\qquad A=-t_1+\lambda t_1+a(1-\lambda),$$

then $A<0$ for large enough t_1. Using Theorems 1 and 2 and Lemma 4, we obtain

$$
\begin{aligned}
(61)\qquad & \int_{\mathcal{B}^c}\left[u_t^2-|\nabla u|^2\right]_{t=t_2}d^3x \\
&\le\int_{V_{\mathcal{B}}^{t_2+A}}\left[u_t^2+|\nabla u|^2\right]_{t=t_2}d^3x+\int_{V_{t_2+A}^{t_2}}\left[u_t^2-|\nabla u|^2\right]_{t=t_2}d^3x \\
&\le\int_{V_{\mathcal{B}}^{t_1+A}}\left[u_t^2+|\nabla u|^2\right]_{t=t_1}d^3x+O\left(\frac{1}{t_1+A}\right) \\
&\quad+\int_{V_{t_2+A}^{t_2}}\left[u_t^2-|\nabla u|^2\right]_{t=t_2}d^3x \\
&=O\left(\frac{1}{t_1}\right)+O\left(\frac{1}{t_1+A}\right)+O\left(\ln\frac{t_2}{t_2+A}\right) \\
&=O\left(\frac{1}{t_1+A}+\ln\frac{t_2}{t_2+A}\right).
\end{aligned}
$$

Relation (61) indicates the time dependence of the difference between kinetic and potential energy.

If we choose $t_1=t_2^{1/2}$ and expand the function

$$
\ln\left[1+\frac{(1-\lambda)\left(\sqrt{t_2}-a\right)}{t_2-(1-\lambda)\left(\sqrt{t_2}-a\right)}\right]
$$

in power series as $t_2\to\infty$, we obtain (59), which proves the theorem.

REFERENCES

[1] D. G. COSTA, *On partition of energy for uniformly propagative systems*, J. Math. Anal. Appl. (1977), pp. 56–62.

[2] G. DASSIOS, *Equipartition of energy for Maxwell's equations*, Quart. Appl. Math., 37 (1980), pp. 465–469.

[3] ———, *Equipartition of energy in elastic wave propation*, Mech. Res. Comm., 6 (1979), pp. 45–50.

[4] G. DASSIOS AND E. GALANIS, *Asymptotic equipartition of kinetic and strain energy for elastic waves in anisotropic media*, Quart. Appl. Math., 38 (1980), pp. 121–128.

[5] G. DASSIOS, *Energy theorems for magnetoelastic waves in a perfectly conducting medium*, Quart. Appl. Math., 39 (1982), pp. 479–490.

[6] R. S. DUFFIN, *Equipartition of energy in wave motion*, J. Math. Anal. Appl., 32 (1970), pp. 386–391.

[7] F. G. FRIEDLANDER, *On the radiation field of pulse solutions of the wave equation*, *Part* I, Proc. Roy. Soc. London Ser. A, 269 (1962), pp. 53–65. *Part* II, 279 (1964), pp. 386–394.

[8] J. A. GOLDSTEIN, *An asymptotic property of solutions of wave equations*, Proc. Amer. Math. Soc., 23 (1969), pp. 359–363.

[9] J. A. GOLDSTEIN AND J. T. SANDEFUR, *Abstract equipartition of energy theorems*, J. Math. Anal. Appl., 67 (1979), pp. 58–74.

[10] H. A. LEVINE, *An equipartition of energy theorem for weak solutions of evolutionary equations in Hilbert space: The Lagrange identity method*, J. Differential Equations, 24 (1977), pp. 197–210.

[11] C. S. MORAWETZ, *The decay of solutions of the exterior initial-boundary value problem for the wave equation*, Comm. Pure Appl. Math., XIV (1961), pp. 561–568.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0007 $01.25/0

SINGULAR PROBLEMS IN THE THEORY OF STRESS-ASSISTED DIFFUSION*

VASILIOS ALEXIADES† AND ELIAS C. AIFANTIS‡

Abstract. A recently developed stress-assisted diffusion theory is further substantiated by establishing the well-posedness of relevant (degenerate parabolic) initial-boundary value problems for a plane with a slit.

1. Introduction. The theory of stress-assisted diffusion proposed by Aifantis (a recent review is provided in [1]) is based on the flux expressions

$$\mathbf{j}=(-D\mathbf{1}+N\sigma\mathbf{1}+K\mathbf{S})\nabla\rho+(L+M\rho)\nabla\sigma, \tag{1.1}$$

with $\mathbf{j}$ denoting the flux of diffusing species, ρ the concentration, σ the trace of the stress tensor $\mathbf{S}$, and D,K,L,M,N positive constants. If $N=K=L=0$, (1.1) becomes identical to Cottrell's postulate of stress-assisted diffusion, if $L=M=0$ it specializes to the flux expression proposed by Flynn, while in the absence of stress (1.1) reduces to the classical Fick's first law of diffusion. Equation (1.1) is a consequence of a differential equation expressing conservation of momentum for the diffusing species. Conservation of mass is expressed by

$$\frac{\partial\rho}{\partial t}+\operatorname{div}\mathbf{j}=f, \tag{1.2}$$

with f representing sources (or sinks) due to chemical reactions, and in the absence of body forces and inertia effects the stress field $\mathbf{S}$ is restricted to satisfy the equations of equilibrium

$$\operatorname{div}\mathbf{S}=0.$$

It is often assumed that the effect of shear stress on diffusion is negligible and that $\mathbf{S}$ is the solution of an elastic problem. It then follows that $K=0$ in (1.1) and that σ is harmonic,

$$\Delta\sigma=0.$$

Under these conditions, substitution of (1.1) into (1.2) yields the differential equation

$$\frac{\partial\rho}{\partial t}=(D+N\sigma)\Delta\rho-\hat{M}\nabla\sigma\cdot\nabla\rho+f,\qquad \hat{M}=M-N\geq 0. \tag{1.3}$$

Equation (1.3) corrects the stress-assisted diffusion equation of Cottrell since it allows the diffusivity to vary linearly with σ. Values of the parameter N have experimentally been determined for certain diffusion systems [1] and the molecular calculations of Aifantis [2] are consistent with this determination. Equation (1.3) with $N=0$, i.e. Cottrell's equation, is often employed by materials scientists to model diffusion controlled processes associated with metallurgical phenomena such as embrittlement, stress corrosion cracking and precipitation in dislocations. Recently, it was shown by Unger and Aifantis [10] and Unger, Gerberich, and Aifantis [11] that predictive models for embrittlement and stress corrosion cracking which are in accord with experiment

*Received by the editors January 15, 1982, and in revised form July 15, 1982.

†Department of Mathematics, The University of Tennessee, Knoxville, Tennessee 37996.

‡Department of Mechanical Engineering and Engineering Mechanics, Michigan Technological University, Houghton, Michigan 49931, and Corrosion Center and Mathematics Institute, University of Minnesota, Minneapolis, Minnesota 55455.

can be obtained by allowing $N \neq 0$ in (1.3). Indeed, as pointed out by Aifantis [4] the equilibrium solution of (1.3) (with $f \equiv 0$) i.e. the solution for which the flux vanishes:

$$\rho = \text{const.} \left(1 + \frac{N}{D}\sigma\right)^{M/N}, \tag{1.4}$$

in conjunction with the singular representation of σ predicted by linear fracture mechanics:

$$\sigma = \frac{\text{const.}}{\sqrt{r}} \cos\frac{\theta}{2}, \tag{1.5}$$

(with r and θ being polar coordinates measured from the crack tip) leads to failure criteria that have experimentally been verified for embrittlement and stress corrosion situations. It is noted that the equilibrium solution of Cottrell's theory can be obtained from (1.4) by letting $N \to 0$ and using Euler's identity. This solution has an exponential form:

$$\rho = \text{const.} \exp\left\{\frac{M}{D}\sigma\right\}, \tag{1.6}$$

but does not lead to failure criteria analogous to those predicted by (1.4).[1]

As is recently reviewed by Aifantis [3] and also detailed by Hill [7] and Unger and Aifantis [10], the equilibrium solution (1.4) with σ as in (1.5) motivates the search for more general plane steady-state solutions of (1.3) by means of the transformation

$$\rho = \rho(x,y) = \hat{\rho}(\xi,\eta),$$

where $\xi = D + N\sigma$ and η is the harmonic conjugate of ξ. Then $\hat{\rho}(\xi,\eta)$ satisfies (when $f \equiv 0$)

$$\xi(\hat{\rho}_{\xi\xi} + \hat{\rho}_{\eta\eta}) + \left(1 - \frac{M}{N}\right)\hat{\rho}_{\xi} = 0,$$

which, with $\alpha \equiv 1 - \frac{M}{N}$, is the equation of generalized axially symmetric potential theory (GASPT):

$$\Delta\hat{\rho} + \frac{\alpha}{\xi}\hat{\rho}_{\xi} = 0.$$

This equation has been studied extensively by Weinstein [12] and his students. Thus, general separable solutions can easily be found (see [10]). Unfortunately, the above change of variables, which leads to the construction of explicit steady-state solutions of (1.3), does not work in general for the transient problem.

The object of this paper is to establish the well-posedness of a general initial-boundary value problem for (1.3) in a plane domain with a crack, at the tip of which σ may becomed infinite. The resulting singularity in the coefficients of (1.3) creates a singular parabolic equation, which is what makes the problem mathematically interesting. The precise problem is stated in §2 and it is transformed into a more convenient form there. In §3, the problem is reformulated in suitably defined Sobolev type spaces and the definition of weak solution is given. The existence, uniqueness and continuous

[1] As shown in [4], the solution (1.4) can lead into a power-law relationship between crack velocities and stress-intensity factors. This relationship, being asserted earlier on empirical grounds, describes reasonably well the trends of experiment as reported in [10], [11]. In this connection, it is noted that the structure of Cottrell's solution (1.6) is not amenable to this type of analysis and results analogous to those derived with the use of (1.4) are not possible in this case.

dependence on the data of the weak solution are established in §4 by means of the Lions projection theorem [8], which is the main tool for singular problems (see for example [6], [5]). Our weak formulation is of the "variational" type and thus well suited for modern numerical methods. We close in §5 with some remarks about important particular cases. In our treatment we consider N to be positive, but even the case $N=0$ (Cottrell's equation) can be included in our results as we point out in the last section.

2. Statement of the problem. In most applications involving (1.3) we are interested in solutions valid in the neighborhood of cracks and dislocations where the stress σ is singular and varies as a negative power of the radial coordinate. Thus we consider a bounded domain with a cut, along which the concentration is prescribed. On the outside boundary, the flux $\mathbf{j}\cdot\mathbf{n}$ may be given or, more generally, a convective type (i.e., third kind) boundary condition may be imposed.

To be precise, consider the (x_1,x_2)-plane cut along the negative x_1-axis and let Ω be its intersection with a bounded domain containing the origin. Points in Ω will be denoted by $x=(x_1,x_2)$. The boundary of Ω consists of the closure $\bar{\kappa}$ of the cut κ (an interval of the $(-x_1)$-axis) and the outside boundary $\Sigma=\partial\bar{\Omega}\setminus\bar{\kappa}$, i.e., $\partial\Omega=\bar{\kappa}\cup\Sigma$. We assume

(2.1) Σ is a $\mathcal{C}^1$-smooth curve bounded away from the origin.

The problem may be stated as follows: Find $\rho(x,t)$ satisfying

$$
\begin{aligned}
&\rho_t=(D+N\sigma)\Delta\rho-\hat{M}\nabla\sigma\cdot\nabla\rho+f(x,t) \quad \text{in } Q:=\Omega\times(0,T),\\
&\rho(x,0)=\rho_0(x), \qquad x\in\Omega,\\
&\rho(x,t)=\rho_\kappa(x), \qquad x\in\kappa,\quad 0<t<T,\\
&J+\alpha(x)[\rho-\rho_\infty(x)]=g(x,t), \qquad x\in\Sigma,\quad 0<t<T,
\end{aligned}
\tag{2.2}
$$

where

$$
J:=-(D+N\sigma)\frac{\partial\rho}{\partial n}+(L+M\rho)\frac{\partial\sigma}{\partial n}
$$

denotes the flux through Σ, $\mathbf{n}$ being the outward unit normal to Σ, $D,N,M,L,\hat{M}=M-N$ are positive physical constants, and $T>0$, $f(x,t)$, $\rho_0(x)$, $\rho_\kappa(x)$, $\alpha(x)$, $\rho_\infty(x)$, $g(x,t)$ constitute the data of the problem, assumptions about which will be made as we proceed. The stress $\sigma(x)$ is a given nonnegative function harmonic in Ω such that

$$
\sigma(x)\big|_\kappa=0, \qquad \sigma(x)=O\big(|x|^{-\gamma}\big) \quad \text{as } |x|\to 0 \text{ with } \gamma<1,
\tag{2.3}
$$

as, for instance, in (1.5).

Let us rewrite the problem in a form more convenient for the analysis. We set

$$
m(x):=(D+N_\sigma(x))^{-1},
\tag{2.4}
$$

$$
\mu(x):=\sqrt{q(q+1)}\,\frac{|\nabla m(x)|}{m(x)}, \qquad q:=\frac{\hat{M}}{2N}=\frac{M-N}{2N}\geq 0,
\tag{2.5}
$$

and note that

$$
0\leq m(x)\leq\frac{1}{D}, \qquad x\in\bar{\Omega},
$$

and that there exists a constant C_1 such that

$$
m(x)\leq C_1\mu(x)^2, \qquad x\in\Omega.
\tag{2.6}
$$

Now we introduce the new unknown

$$v(x,t) := m(x)^q \rho(x,t), \tag{2.7}$$

for which the problem takes the form

$$\begin{aligned} & m(x)v_t = \Delta v - \mu(x)^2 v + m(x)^{q+1} f(x,t) \quad \text{in } Q, \\ & v(x,0) = v_0(x) := m(x)^q \rho_0(x), \qquad x \in \Omega, \\ & v(x,t) = v_1(x) := D^{-q}\rho_\kappa(x), \qquad x \in \kappa, \quad 0<t<T \\ & \frac{\partial v}{\partial n} + p(x)v = v_2(x,t), \qquad x \in \Sigma, \quad 0<t<T. \end{aligned} \tag{2.8}$$

where

$$p(x) := -m(x)\left[\alpha(x) + \frac{M+N}{2}\frac{\partial \sigma}{\partial n}\right], \qquad x \in \Sigma, \tag{2.9}$$

$$v_2(x,t) := m(x)^{q+1}\left[g(x,t) - L\frac{\partial \sigma}{\partial n} + \alpha(x)\rho_\infty(x)\right], \qquad x \in \Sigma, \quad 0<t<T. \tag{2.10}$$

In the next section, a weak formulation in appropriate weighted Sobolev spaces will be obtained. For that we need the solution to vanish along the crack κ and in order to achieve it we must be able to extend the data $v_1(x)$ inside Ω. We assume

$$\rho_\kappa \in \mathcal{C}^1(\bar{\kappa}) \text{ and, without loss of generality, } \rho_\kappa(0) = 0, \tag{2.11}$$

(otherwise, consider $\rho - \rho_\kappa(0)$ instead of ρ, which only shifts the data $\rho_0(x)$, $\rho_\kappa(x)$, $\rho_\infty(x)$ and $g(x,t)$ in (2.2)). We extend $\rho_\kappa(x)$ to all of the negative x_1-axis as a $\mathcal{C}^1$-function and then extend it to the whole plane radially, i.e., we define

$$\tilde{\rho}_\kappa(x) \equiv \tilde{\rho}_\kappa(r,\theta) := \rho_\kappa(r), \qquad 0 \le r < \infty, \quad -\pi < \theta \le \pi.$$

Then $\tilde{\rho}_\kappa \in \mathcal{C}^1(\bar{\Omega})$, so also $v_1(x) = D^{-q}\rho_\kappa(x)$ is extended to a function

$$\tilde{v}_1(x) = D^{-q}\tilde{\rho}_\kappa(x), \quad \tilde{v}_1 \in \mathcal{C}^1(\bar{\Omega}), \quad \tilde{v}_1(0) = 0. \tag{2.12}$$

3. Weak formulation. Let $\phi(x,t)$ be a smooth function defined in $\bar{Q}$ such that

$$\phi|_{x \in \kappa} = 0, \qquad \phi|_{t=T} = 0.$$

Multiplying the equation in (2.8) by ϕ and integrating over Q yields

$$\iint_Q \{\nabla v \cdot \nabla \phi + \mu^2 v\phi - mv\phi_t\}\, dx\, dt = \int_0^t \int_\Sigma \phi \frac{\partial v}{\partial n}\, d\Sigma\, dt + \int_\Omega m(x)v_0(x)\phi(x,0)\, dx + \iint_Q m^{q+1} f\phi\, dx\, dt.$$

Now, setting

$$u(x,t) := v(x,t) - \tilde{v}_1(x), \tag{3.1}$$

with $v_1(x)$ as in (2.12), we find that $u(x,t)$ must satisfy

(3.2)

$$\begin{aligned} \iint_Q \{\nabla u \cdot \nabla \phi + \mu^2 u\phi - mu\phi_t\}\, dx\, dt + \int_0^T \int_\Sigma pu\phi\, d\Sigma\, dt \\ = \int_0^T \int_\Sigma u_2 \phi\, d\Sigma\, dt + \int_\Omega mu_0(x)\phi(x,0)\, dx + \iint_Q m^{q+1} f\phi\, dx\, dt \\ - \iint_Q \{\nabla \tilde{v}_1 \cdot \nabla \phi + \mu^2 \tilde{v}_1 \phi\}\, dx\, dt, \end{aligned}$$

where we have set

$$u_0(x) := v_0(x) - \tilde{v}_1(x), \qquad x \in \Omega, \tag{3.3}$$

$$u_2(x,t) := v_2(x,t) - p(x)\tilde{v}_1(x), \qquad x \in \Sigma, \quad 0 < t < T. \tag{3.4}$$

Our definition of weak solution will be based on (3.2), which leads us to consider the following function spaces.

Let $\hat{\mathcal{C}}$ be the set of all $\mathcal{C}^1(\bar{\Omega})$ functions ζ which vanish on κ and such that $\int_\Omega \mu(x)^2 \zeta^2 \, dx < \infty$. Let

$$\mathcal{H} \text{ be the closure of } \hat{\mathcal{C}} \text{ in the norm } \|h\|_{\mathcal{H}} := \left(\int_\Omega \mu(x)^2 |h(x)|^2 dx \right)^{1/2}$$

and

$$\mathcal{W} \text{ be the closure of } \hat{\mathcal{C}} \text{ in the norm } \|w\|_{\mathcal{W}} := \left(\int_\Omega \left\{ \mu(x)^2 |w(x)|^2 + |\nabla w|^2 \right\} dx \right)^{1/2}.$$

Then $\mathcal{H}$ and $\mathcal{W}$ are Hilbert spaces and $\|w\|_{\mathcal{H}} \le \|w\|_{\mathcal{W}}$, $w \in \mathcal{W}$, so that $\mathcal{W}$ is densely and continuously imbedded into $\mathcal{H}$.

Regarding the vanishing of elements of $\mathcal{W}$ on κ we have the following.

THEOREM 3.1. *If $w \in \mathcal{W} \cap \mathcal{C}(\Omega \cup \kappa)$ then $w = 0$ on κ.*

Proof. For convenience in this proof we shall denote points of $\bar{\Omega}$ by (x,y). Fix a point $(x^\circ, 0) \in \kappa$ and consider a rectangle $R := B \times (0,h) \subset \Omega$ with base an interval $\bar{B} := [x^\circ - \delta, x^\circ + \delta] \subset \kappa$, $\delta > 0$, and height $h > 0$. Let $\{w_n\}$ be a sequence in $\hat{\mathcal{C}}$ which converges in $\mathcal{W}$ to w. For any $(x,y) \in R$, we have

$$w_n(x,y) = \int_0^y \frac{\partial}{\partial y} w_n(x,\eta) \, d\eta,$$

whence

$$\iint_R \mu(x,y)^2 |w_n(x,y)|^2 dx\, dy \le \iint_R \mu(x,y)^2 y \left(\int_0^h \left| \frac{\partial}{\partial y} w_n(x,\eta) \right|^2 d\eta \right) dx\, dy.$$

But on $\bar{R}$, μ is a bounded quantity, say $\mu^2 \le C_2$, so

$$\iint_R \mu^2 |w_n|^2 dx\, dy \le C_2 \frac{h^2}{2} \int_B \int_0^h \left| \frac{\partial}{\partial y} w_n \right|^2 d\eta\, dx \le \frac{C_2 h^2}{2} \iint_R |\nabla w_n|^2 dx\, dy.$$

Taking $n \to \infty$, we have

$$\frac{1}{h} \int_0^h \int_B \mu^2 |w|^2 dx\, dy \le \frac{hC_2}{2} \iint_R |\nabla w|^2 dx\, dy. \tag{3.5}$$

Now, by the mean value theorem, the left-hand side of (3.5) equals

$$\int_B \mu(x,y^*)^2 |w(x,y^*)|^2 dx \quad \text{for some } 0 < y^* < h.$$

Letting $h \to 0$, we find

$$\int_B \mu(x,0)^2 |w(x,0)|^2 dx = 0,$$

hence $w(x,0) = 0$, $x \in B$. Q.E.D.

We shall need the following

LEMMA 3.2. *The function $\tilde{v}_1$ of* (2.12) *belongs to the space $\mathcal{W}$.*

Proof. From (2.12), $|\nabla\tilde{v}_1|\in L^2(\Omega)$. Since the only singularity of $\mu(x)$ is at the origin, we shall have $\|\tilde{v}_1\|_{\mathfrak{W}}<\infty$ if $\mu^2|\tilde{v}_1|^2$ is integrable in a neighborhood of the origin. Now, near $x=0$, by (2.5), (2.4) and (2.3),

$$\mu(x)^2\le \text{const.}\, r^{-2\gamma-2} \quad \text{as } r=|x|\to 0,$$

and $\tilde{v}_1(x)=v_1(r)$, so that for small $R>0$ we have (using (2.12))

$$\begin{aligned}\iint_{|x|<R}\mu^2|\tilde{v}_1|^2dx &\le C_3\int_{-\pi}^{\pi}\int_0^R r^{-2\gamma-2}|v_1(r)|^2 r\,dr\,d\theta\\ &= C_4\int_0^R r^{1-2\gamma}\left|\frac{v_1(r)-v_1(0)}{r}\right|^2 dr\\ &\le C_4\sup_{\bar{\kappa}}\left|\frac{d}{dr}v_1(r)\right|^2\int_0^R r^{1-2\gamma}dr\\ &<\infty,\end{aligned}$$

because $\gamma<1$. Q.E.D.

For any Hilbert space $\mathcal{V}$, $L^2(0,T;\mathcal{V})$ will denote the usual Hilbert space with norm

$$\|v\|_{L^2(0,T;\mathcal{V})}:=\left(\int_0^T\|v(\cdot,t)\|_{\mathcal{V}}^2dt\right)^{1/2}<\infty.$$

Now we consider the bilinear form

$$B[u,\phi]:=\iint_Q\left\{\nabla u\cdot\nabla\phi+\mu(x)^2u\phi-m(x)u\phi_t\right\}dx\,dt+\int_0^T\int_\Sigma p(x)u\phi\,d\Sigma\,dt, \tag{3.6}$$

and the linear functional

$$\begin{aligned}\Lambda[\phi]:=&\int_0^T\int_\Sigma u_2(x,t)\phi\,d\Sigma\,dt+\int_\Omega m(x)u_0(x)\phi(x,0)\,dx\\ &+\iint_Q m(x)^{q+1}f(x,t)\phi\,dx\,dt-\iint_Q\left\{\nabla\tilde{v}_1(x)\cdot\nabla\phi+\mu(x)^2\tilde{v}_1(x)\phi\right\}dx\,dt.\end{aligned} \tag{3.7}$$

For $u\in L^2(0,T;\mathfrak{V})$ and $\phi\in L^2(0,T;\mathfrak{V})$ with $\phi_t\in L^2(0,T;\mathcal{H})$, all the integrals in (3.6) and (3.7) will be finite provided the original data satisfy the following assumptions:

$$m^{q+1/2}f\in L^2(Q), \tag{3.8}$$

$$m^{q+1/2}\rho_0\in L^2(\Omega), \tag{3.9}$$

$$\alpha\in L^\infty(\Sigma) \quad\text{and}\quad \alpha(x)\le-\frac{M+N}{2}\frac{\partial\sigma}{\partial n} \quad \text{a.e. on } \Sigma, \tag{3.10}$$

$$\rho_\infty\in L^2(\Sigma),\ g\in L^2(\Sigma\times(0,T)). \tag{3.11}$$

Remark 1. Assumptions (3.8) and (3.9) are weaker than $f\in L^2(Q)$ and $\rho_0\in L^2(\Omega)$, respectively, when $0<\gamma<1$, but equivalent to these when $\gamma\le 0$.

Remark 2. Assumption (3.10) guarantees $p\in L^\infty(\Sigma)$ and $p(x)\ge 0$ a.e. on Σ which is needed for existence and uniqueness of the solution (see also §5).

Remark 3. The finiteness of $\iint_Q m(x)u\phi_t\,dx\,dt$ follows from (2.6).

Remark 4. $\phi\in L^2(0,T;\mathfrak{V})$ and $\phi_t\in L^2(0,T;\mathcal{H})$ imply [9, p. 19] that $t\mapsto\phi(\cdot,t)$ is continuous on $[0,T]$ so that $\phi(\cdot,0)$ and $\phi(\cdot,T)$ make sense in $\mathcal{H}$. Moreover, $\phi(\cdot,t)\in\mathfrak{V}$

implies (since $\mu(x)^2$ is bounded and nonvanishing near Σ) that the trace of ϕ on Σ exists in $L^2(\Sigma)$. Similarly, the trace of u on Σ exists in $L^2(\Sigma)$.

Now we are in a position to define our concept of weak solution.

DEFINITION. By a *weak solution* of (the transformed) problem (2.8) we mean a function v of the form

$$v(x,t)=u(x,t)+\tilde{v}_1(x) \tag{3.12}$$

for some function $u\in L^2(0,T;\mathfrak{V})$ satisfying

$$B[u,\phi]=\Lambda[\phi] \tag{3.13}$$

for every $\phi\in L^2(0,T;\mathfrak{V})$ with $\phi_t\in L^2(0,T;\mathfrak{H})$ and $\phi(\cdot,T)=0$ in $\mathfrak{H}$.

The original unknown $\rho(x,t)$ is given in terms of $v(x,t)$ by

$$\rho(x,t)=(D+N\sigma(x))^q v(x,t), \qquad q=\frac{\tilde{M}}{2N}. \tag{3.14}$$

Remark 5. Clearly, a classical solution is a weak solution, and conversely, a sufficiently regular weak solution is also classical.

Remark 6. Our weak solution is of "variational" type in a sense opposite to "operational" and by analogy to time-independent problems. In fact, in the steady-state case the weak solution is an extremum of the functional

$$\int_\Omega\left\{\frac{1}{2}\left(|\nabla u|^2+\mu^2u^2\right)+\nabla\tilde{v}_1\cdot\nabla u+\mu^2\tilde{v}_1u-m^{q+1}fu\right\}dx+\int_\Sigma\left\{\frac{1}{2}pu^2-u_2u\right\}d\Sigma.$$

In any case, (3.13) is suitable for Galerkin-type numerical methods.

4. Existence and uniqueness.

THEOREM 4.1 (well-posedness). *Under the assumptions* (2.1), (2.3), (2.11) *and* (3.8)–(3.11) *on the data, problem* (2.8) *has unique weak solution* $v(x,t)$ *which also satisfies the estimate*

$$\left(\int_0^T\left\{\|v\|_{\mathfrak{V}}^2+\left\|\sqrt{p}\,v\right\|_{L^2(\Sigma)}\right\}dt\right)^{1/2}\le\sqrt{C_1}\,\|m^{q+1/2}f\|_{L^2(Q)}+\sqrt{2}\left\|\sqrt{m}\,v_0\right\|_{L^2(\Omega)}$$
$$+C_5\|\tilde{v}_1\|_{\mathfrak{V}}+C_6\left\|\sqrt{p}\,\tilde{v}_1\right\|_{L^2(\Sigma)}+\|v_2\|_{L^2(\Sigma\times(0,T))}, \tag{4.1}$$

with the constants C_5, C_6 *depending only on* C_1 *of* (2.6), $\|p\|_{L^\infty(\Sigma)}$ *and* T.

Proof. Existence of a solution will be obtained by means of the Lions generalization of the projection theorem (Lions [8, p. 37]) with the following choice of spaces:

$$\mathfrak{F}:=L^2(0,T;W)$$

with

$$\|u\|_{\mathfrak{F}}:=\left(\int_0^T\|u(\cdot,t)\|^2dt+\int_0^T\int_\Sigma p(x)|u|^2d\Sigma\,dt\right)^{1/2},$$
$$\Phi:=\{\phi\in f:\ \phi_t\in L^2(0,T;\mathfrak{H}),\ \phi(\cdot,T)=0\text{ in }\mathfrak{H}\}$$

with

$$\|\phi\|_\Phi:=\left(\|\phi\|_{\mathfrak{F}}^2+\frac{1}{2}\int_\Omega m(x)\phi(x,0)^2dx\right)^{1/2}.$$

Then $\mathfrak{F}$ is a Hilbert space, Φ is an (incomplete) inner product space, $\Phi\subset\mathfrak{F}$ and $\|\phi\|_{\mathfrak{F}}\le\|\phi\|_\Phi$, $\phi\in\Phi$. The bilinear functional $B[u,\phi]$ defined on $\mathfrak{F}\times\Phi$ by (3.16) and the

linear functional $\Lambda[\phi]$ defined on Φ by (3.11) satisfy the following:

CLAIM 1. *$B[\cdot,\cdot]$ is coercive on $\Phi\times\Phi$.*

In fact, one immediately sees that $B[\phi,\phi]=\|\phi\|_{\Phi}^2$, $\phi\in\Phi$.

CLAIM 2. *$B[\cdot,\phi]$ is bounded on $\mathcal{F}$ for each $\phi\in\Phi$.*

Indeed, using Holder's inequality and (2.10) we find for each $\phi\in\Phi$

$$|B[u,\phi]|\leq (\text{constant depending on } \phi)\|u\|_{\mathcal{F}}, \qquad u\in f.$$

CLAIM 3. *$\Lambda[\cdot]$ is bounded on Φ.*

Indeed,

$$|\Lambda[\phi]|\leq \Big\{\|u_2\|_{L^2(\Sigma\times(0,T))}+\sqrt{2}\,\big\|\sqrt{m}\,u_0\big\|_{L^2(\Omega)} + \sqrt{C_1}\,\|m^{q+1/2}f\|_{L^2(Q)}+2\sqrt{T}\,\|\tilde{v}_1\|_{\mathcal{W}}\Big\}\|\phi\|_{\Phi}, \qquad \phi\in\Phi.$$

Thus, by the Lions theorem referred to above, there exists $u\in\mathcal{F}$ satisfying $B[u,\phi]=\Lambda[\phi]$, $\phi\in\Phi$, and such that $\|u\|_{\mathcal{F}}\leq\|\Lambda\|$, the operator norm of Λ, i.e.

$$\|u\|_{\mathcal{F}}\leq\|u_2\|_{L^2(\Sigma\times(0,T))}+\sqrt{2}\,\big\|\sqrt{m}\,u_0\big\|_{L^2(\Omega)}+\sqrt{C_1}\,\|m^{q+1/2}f\|_{L^2(Q)}+2\sqrt{T}\,\|\tilde{v}_1\|_{\mathcal{W}},$$

which implies (4.1). This estimate however is not an a priori bound, so uniqueness has not been proved yet.

To establish the uniqueness of the solution, let v_1 and v_2 be any two solutions. Then $v:=v_1-v_2$ is a weak solution of the homogeneous problem (2.8), i.e. $u\equiv v\in\mathcal{F}$ satisfies

$$B[u,\phi]=0, \qquad \phi\in\Phi. \tag{4.2}$$

For any $0<\tau<T$, let $Q^{\tau}:=\Omega\times(0,\tau)$ and consider the function

$$\psi(x,t):=\begin{cases}\displaystyle\int_t^{\tau}u(x,s)\,dx & \text{for } 0\leq t\leq\tau,\\ 0 & \text{for } \tau\leq t\leq T,\end{cases}$$

which also belongs to $\mathcal{F}$. Then $\psi_t=-u\in\mathcal{F}\subset L^2(0,T;\mathcal{H})$ and $\psi(\cdot,t)=0$, so $\psi\in\Phi$. In (4.2) we take $\phi=\psi$ and $u=-\psi_t$ to find (after some integrations by parts and because $\psi(x,\tau)=0$, $x\in\Omega$ implies $\nabla\psi(x,\tau)=0$)

$$\frac{1}{2}\int_{\Omega}|\nabla\psi(x,0)|^2dx+\frac{1}{2}\int_{\Omega}\mu(x)^2\psi(x,0)^2dx+\iint_{Q}m\psi_t^2\,dx\,dt+\frac{1}{2}\int_{\Sigma}p(x)\psi(x,0)^2d\Sigma=0$$

Thus, $m(x)\psi_t^2=0$ a.e. in Q^{τ}, i.e. $u(x,t)=0$ a.e. in $\Omega\times(0,\tau)$, any $0<\tau<T$. Q.E.D.

5. Concluding remarks.

The problem with prescribed flux. The boundary condition on Σ in (2.2) reduces to that of prescribed flux wherever $\alpha(x)=0$. In view of our condition (3.10), this is allowed whenever $\partial\sigma(x)/\partial n|_{\Sigma}\leq 0$ but not otherwise. Thus, for the typical stress given by (1.5), the flux could be prescribed at any point of Σ where the tangent line does not pass through the origin. Notice moreover that in the case $M=N=0$ (as in Fick's law), condition (3.10) reduces to the classical one, namely, $\alpha(x)\leq 0$.

Nonsingular stress. We have allowed the stress $\sigma(x)$ to be singular at the crack tip of order $|x|^{-\gamma}$ with $\gamma<1$ (see (2.3)). Thus, the case of nonsingular stress is included. The problem will be uniformly parabolic in Ω if $\gamma\leq -1$, but still singular in the lower order term if $-1<\gamma\leq 0$.

Steady state problems. By taking $\rho(x,t)$ (as well as $v(x,t)$ and $u(x,t)$) independent of t and omitting the initial condition in (2.2) (and (2.8)), we obtain the steady-state problem. Thus, its well-posedness has also been established.

Cottrell's equation. Even though we have assumed $N>0$ in our treatment, the reformulated problem (2.8) (and consequently all the results) has the same form even for $N=0$, except that now $v(x,t)$ in (2.7) must be defined as

$$v(x,t):=e^{-\tilde{q}}\rho(x,t) \quad \text{with } \tilde{q}:=\frac{M}{2D}.$$

Indeed, as $N\to 0$, $m(x)$ becomes simply $1/D$ and (see (2.5))

$$\mu(x)=\sqrt{q(q+1)}\,\frac{|\nabla m|}{m}\equiv\sqrt{\frac{M-N}{2}\left(\frac{M-N}{2}+N\right)}\,\frac{|\nabla\sigma|}{D+N\sigma}$$

becomes $\tilde{q}|\nabla\sigma|$. With these redefinitions of $q, m,$ and $\mu,$ everything in §3 and 4 is still valid and we obtain the existence, uniqueness and continuous dependence on the data of the weak solution $v(x,t)$ of (2.8). Then, the solution of the problem for Cottrell's equation is given by

$$\rho(x,t)=e^{\tilde{q}}v(x,t), \qquad \tilde{q}=\frac{M}{2D},$$

instead of (3.14).

Acknowledgment. The support of the Solid Mechanics Program of the National Science Foundation and the Corrosion Center of the University of Minnesota is gratefully acknowledged.

REFERENCES

[1] E. C. AIFANTIS, *On the problem of diffusion in solids*, Acta Mechanica, 37 (1980), pp. 265–296.

[2] _____, *Comments on the calculation of the formation volume of vacancies in solids*, Phys. Rev., B, 19 (1979), pp. 6622–6624.

[3] _____, *The mechanics of diffusion in solids*, TAM Report 440, Univ. Illinois, Urbana, March 1980.

[4] _____, *Elementary physicochemical degradation processes*, in Proc. International Symposium on the Mechanics of Behavior of Structured Media, A.P.S. Selvadurai, ed., Carleton University, Ottawa, Canada, May 18–21, 1981, pp. 301–317.

[5] V. ALEXIADES, *A singular parabolic initial-boundary value problem in a noncylindrical domain*, this Journal, 11 (1980), pp. 348–357.

[6] R. W. CARROLL, AND R. E. SHOWALTER, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.

[7] J. M. HILL, *Plane steady solutions for stress-assisted diffusion*, Mech. Res. Comm., 6 (1979), pp. 147–150.

[8] J. L. LIONS, *Equations différentielles opérationelles et problèmes aux limites*, Springer-Verlag, Berlin, 1961.

[9] J. L. LIONS, AND E. MAGENES, *Nonhomogeneous Boundary Value Problems*, vol. I, Springer-Verlag, New York, 1972.

[10] D. J. UNGER AND E. C. AIFANTIS, *On the theory of stress-assisted diffusion-II*, Acta Mechanica, in press.

[11] D. J. UNGER, W. W. GERBERICH, AND E. C. AIFANTIS, *Further remarks on the implications of steady-state stress-assisted diffusion on environmental cracking*, Scripta Metallurgica, 16 (1982), pp. 1059–1064.

[12] A. WEINSTEIN, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–38.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0008 $01.25/0

A TWO-DIMENSIONAL DIRICHLET PROBLEM WITH AN EXPONENTIAL NONLINEARITY*

J. L. MOSELEY†

Abstract. We consider the two-dimensional nonlinear Dirichlet problem

$$\begin{aligned} -\Delta u &= \lambda(e^u + \psi e^{-u}), && y \in \Omega, \\ u &= \phi, && y \in \partial\Omega, \end{aligned}$$

where $y=(y_1,y_2)$, Δ is the Laplacian operator, Ω is a simply connected region bounded by a smooth closed Jordan curve, the boundary data is continuous, ψ is analytic in Ω and continuous on $\bar{\Omega}$, and λ is positive. Our concern is with obtaining a large norm (second) solution for λ tending to 0_+. This is accomplished by obtaining a third-order asymptotic solution which is used as a first approximation for a modified Newton's method. Additionally, we obtain three solutions as $\lambda \to 0^+$ if ψ is a negative constant.

AMS-MOS subject classification (1980). Primary 35J25, 35J60, 35J65.

Key words. elliptic equation, nonlinear boundary value problem

Introduction. We first consider the two-dimensional nonlinear Dirichlet problem

$$\text{(P)} \qquad \begin{aligned} -\Delta u &= \lambda e^u + \nu e^{-u}, && y \in \Omega, \\ u(y) &= \phi(y), && y \in \partial\Omega, \end{aligned}$$

where $y=(y_1,y_2)$, Δ is the Laplacian operator, Ω is a simply connected region bounded by a Jordan curve, the boundary data $\phi(y)$ is continuous on $\partial\Omega$, and ν and λ are real.

There are two noteworthy cases depending on the signs of the parameters λ and ν:

Case A. $\lambda \le 0$, $\nu \ge 0$.

Case B. $\lambda > 0$. By letting $u(y) = -v(y)$, $\lambda = -\nu_1$, and $\nu = -\lambda_1$ we obtain the equivalent problem:

$$\text{(P}') \qquad \begin{aligned} -\Delta v &= \lambda_1 e^v + \nu_1 e^{-v}, && y \in \Omega, \\ v &= -\phi, && y \in \partial\Omega, \end{aligned}$$

which is also of type (P). The case $\lambda \le 0$, $\nu < 0$ for (P) is equivalent to the case $\lambda_1 > 0$, $\nu_1 \ge 0$ for (P′) and is therefore covered by Case B.

It is well known that there exists a unique solution for Case A ([4, p. 323, 372]). In [11] a Newton type method is given for Case A which converges quadratically for any choice of the initial data.

Case B is a special case of

$$\text{(P1)} \qquad \begin{aligned} -\Delta u &= \lambda(e^u + \psi e^{-u}), && y \in \Omega, \\ u &= \phi, && y \in \partial\Omega, \end{aligned}$$

where ψ is a real analytic function of y on Ω and continuous on $\bar{\Omega}$ and the assumption $\lambda > 0$ is part of the hypothesis.

The purpose of this paper is to extend the results of [9] on the large norm (second) solution of (P1) for $\psi \equiv 0$ to the more general case. The addition of the term ψe^{-u} yields the possibility of better steady state models for physical phenomena having exponential nonlinearities, especially nonlinear diffusion processes (e.g., chemical reactors [1]; also see [9] for other possible applications).

*Received by the editors June 9, 1981, and in final form September 21, 1982.

†Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506.

In §1 we review previous results and summarize the extensions contained in this paper. Sections 2 through 6 give the requisite development for the proof of these extensions. Section 7 gives a third solution in the special case where ψ is a negative constant.

1. Summary of results. It is well known that a solution of (P1) is always obtainable for λ sufficiently small [4, p. 373]. Also, the assumption that ψ is analytic means that all solutions of $-\Delta u=\lambda(e^u+\psi e^{-u})$ are analytic in Ω. If in addition the boundary $\partial\Omega$ and the boundary data ϕ are analytic, all solutions to the boundary value problem are analytic in $\bar{\Omega}$ [8]; however, we will consider weaker sufficient conditions to obtain classical solutions, $u\in C^2(\Omega)\cap C(\bar{\Omega})$.

To obtain an equivalent problem with zero boundary data, we let u_0 be the solution of the associated harmonic problem ($\lambda=0$) and rewrite (P1) as

$$\text{(P2)}\qquad \begin{aligned} -\Delta u_1&=\lambda\left(e^{u_0+u_1}+\psi e^{-(u_0+u_1)}\right), && y\in\Omega,\\ u_1&=0, && y\in\Omega, \end{aligned}$$

where $u_1(y)=u(y)-u_0(y)$.

If $\psi(y)\geq 0$ for $y\in\bar{\Omega}$ then all solutions of (P2) are superharmonic and hence positive. If $0\leq\psi(y)<e^{2u_0}$ we can apply the technique of Keller and Cohen [7] for points λ in the "spectrum" (the set of all $\lambda>0$ such that (P2) has a solution) to obtain the "minimal" solution. Keller and Cohen's results also show that the spectrum is a finite interval with least upper bound λ^* satisfying $\lambda^*\leq\mu_1$ where μ_1 is the least eigenvalue of the linear problem

$$\text{(L1)}\qquad \begin{aligned} -\Delta v&=\mu e^{u_0}v, && y\in\Omega,\\ v&=0, && y\in\partial\Omega. \end{aligned}$$

Bandle's results [2] show in the case $0\leq\psi(y)<e^{2\check{u}_0}$ where $\check{u}_0=\min_{y\in\Omega}(u_0(y))$ that for some $\varepsilon>0$

$$\lambda^*\geq K_0\frac{4\pi}{A}+\varepsilon$$

where $K_0=\max_{m\geq 0}(m(e^{m+\hat{u}_0}+\hat{\psi}e^{-m-\check{u}_0})^{-1})$, $\hat{u}_0=\max_{y\in\bar{\Omega}}(u_0(y))$, $\hat{\psi}=\max_{y\in\bar{\Omega}}(\psi(y))$, and A is the area of Ω. Crandall and Rabinowitz's results [5] show that, for $0\leq\psi(y)<e^{2u_0}$, λ^* is in the spectrum and there are two solutions for every $\lambda\in(0,\lambda^*)$.

For the case $\psi\equiv 0$, $u_0\equiv 0$ ($\phi\equiv 0$ for (P1)) and $\partial\Omega$ is a circle, the two solutions of (P2) are given in [2] (and [9]). For $\psi\equiv 0$, $u_0\equiv 0$ ($\phi\equiv 0$) and $\partial\Omega$ the smooth boundary of a simply connected region, Weston [13] developed an asymptotic approximation of the "large norm" (second solution) by using the general integral of $-\Delta u=\lambda e^u$; however, additional implicit constraints on the domain were required. For λ sufficiently small, Weston also showed that if this asymptotic solution is used as a first approximation in an appropriate modified Newton's iteration scheme, then an exact large norm solution is generated provided that the asymptotic solution is taken to order greater than or equal to three.

As no difficulty will arise, we will consider a region $\Omega\subseteq\mathbb{R}^2$ to also be a subset of $\mathbb{C}$ throughout this paper, using $y=(y_1,y_2)\in\Omega\subset\mathbb{R}^2$ and $w=y_1+iy_2\in\Omega\subseteq\mathbb{C}$. When the region is the unit disc U, we will use $x=(x_1,x_2)\in U\subseteq\mathbb{R}^2$ and $z=x_1+ix_2\in\mathbb{C}$. For any region $\Omega\subseteq\mathbb{C}$ with smooth boundary, we let $H(\Omega)$ be the holomorphic functions on Ω, $A(\Omega)$ be the functions in $H(\Omega)$ which are continuous on $\bar{\Omega}$, $A_\alpha(\Omega)$ be the functions in $A(\Omega)$ that satisfy a Lipschitz condition (as a function of arclength) of order α on $\partial\Omega$, $0<\alpha\leq 1$, and $A_N(\Omega)$ be the functions in $A(\Omega)$ which are nonzero in $\bar{\Omega}$. Additionally,

H^p, $0<p\leq\infty$ are the usual Hardy spaces on U and we let H_N be the functions in H which are nonzero in U. If the region is not specified, it is assumed to be U.

In [9], one of the three constraints of Weston is removed and the other two, which are given in terms of the conformal map of the unit disc U onto Ω, are examined and extended to the case of arbitrary boundary data. As in [9], to handle arbitrary boundary data, we let $h(w)=u_0(y)+iv_0(y)$ where u_0 is the solution of the associated harmonic problem ($\lambda=0$) and v_0 is its conjugate harmonic function. Next we let f_Ω be the conformal map of U onto Ω and $f_\phi(z)=\exp\{(1/2)h(f_\Omega(z))\}$. The function f_ϕ characterizes the boundary data whereas f'_Ω characterizes Ω (except for translations).

We refer to [9] and [10] for the relationships between the smoothness of $\partial\Omega$ and the continuity and smoothness of ϕ, and the behavior of f'_Ω and f_ϕ on the boundary of U. In this paper we assume that $f'_\Omega, f_\phi\in A_N(U)$ so that the first constraint of Weston (extended to arbitrary boundary data) is satisfied. Hence (P2) is equivalent to its conformal transplantation

$$\text{(P3)}\qquad \begin{aligned} -\Delta u_1&=\lambda|f_\phi f'_\Omega|^2\left(e^{u_1}+\psi|f_\phi|^{-4}e^{-u_1}\right), && |z|<1,\\ u_1&=0, && |z|=1, \end{aligned}$$

where

$$u_1=u(f_\Omega(z))-\ln|f_\phi(z)|^2.$$

We refer to f where $f'=f_\phi f'_\Omega$ as an associated map for (P1); recall from [9] that it is not unique but depends on the choice of conformal map f_Ω; and note that the form of (P3) is independent of this choice. Then, as in [9], we choose a normalized associated map f_N such that $f_N(0)=0$, $f'_N(0)>0$, and $f''_N(0)=0$, as well as $f'_N\in A_N(U)$. The second condition of Weston (extended to f_N) becomes

$$|f'''_N(0)|\neq 2|f'_N(0)|.$$

Hence, it will suffice to consider

$$\text{(P4)}\qquad \begin{aligned} &\text{PDE} && -\Delta u=\lambda|f'(z)|^2(e^u+\sigma(x)e^{-u}), && |x|<1,\\ &\text{BC} && u=0, && |x|=1, \end{aligned}$$

where f satisfies

(A) $f'\in A_N$, $f(0)=0$, $f'(0)>0$, $f''(0)=0$, $|f'''(0)|\neq 2|f'(0)|$

and σ satisfies

(B) σ is real analytic in Ω and continuous on $\bar{\Omega}$.

As there is no general integral of the PDE given in (P4), we develop a third order asymptotic solution which approximates a large norm solution of the PDE as well as approximating the BC. To do this we make use of the large norm asymptotic solution for the case $\sigma(x)\equiv 0$. Next we develop a modified Newton's iteration scheme which will converge for λ sufficiently small to an exact large norm solution of an equivalent integral equation formulation of (P4) when the asymptotic solution is used as a first approximation. From this we may obtain an exact large norm (second) solution of (P4) and hence (P1).

We also show (by the technique given in the introduction) that if ψ is a negative constant then there are three solutions to (P2) (hence three solutions to (P1)) for λ sufficiently small, the extra "negative" large norm solution being obtainable from the "positive" large norm solution of an equivalent problem.

Thus the main results of this paper are summarized in the following:

THEOREM 1. *If f satisfies* (A) *and σ satisfies* (B), *then an asymptotic solution of* (P4) *of order three can be obtained which will generate, for λ sufficiently small, an exact large norm (second) solution via a modified Newton's iteration scheme for an equivalent integral equation. The asymptotic solution is given by*

$$u_A = u_e(z;\lambda) + H(z;\lambda)$$

where

$$e^{-u_e(z;\lambda)/2} = \frac{|z|^2 + \frac{\lambda}{8}\left|z\int^z [G(\hat{z};\lambda)]^2 \frac{f_n'(\hat{z})}{\hat{z}^2}\,d\hat{z}\right|^2}{|G(z;\lambda)|^2}, \tag{1.2}$$

$$G(z:\lambda) = 1 + \lambda G_1(z) + \lambda^2 G_2(z),$$

$$H(z:\lambda) = \lambda H_1(z) + \lambda^2 H_2(z);$$

the functions G_i and H_i being given in §2.

THEOREM 2. *Let the normalized associated map f_N of* (P1) *satisfy* (A). *Furthermore, let g_Ω be the inverse of f_Ω, u_A be the asymptotic solution of* (P4) *for $\sigma = \psi|f_\phi|^{-4}$, and u_F be the exact solution generated by the Newton's method in Theorem* 1. *Then a large norm asymptotic solution of* (P1) *is given by $u_A(g_\Omega(w)) + \ln|f_\phi(g_\Omega(w))|$ and an exact solution is given by $u_F(g_\Omega(w)) + \ln|f_\phi(g_\Omega(w))|$.*

2. An asymptotic solution for (P4). We choose our asymptotic solution to be of the form:

$$u = u_e(z:\lambda) + H(z:\lambda) \tag{2.1}$$

where

$$H(z:\lambda) = \lambda H_1(z) + \lambda^2 H_2(z) + \cdots + \lambda^{n-1} H_{n-1}(z)$$

and $u_e(z:\lambda)$ is the large norm solution for $-\Delta u = \lambda|f'(z)|^2 e^u$ given by (1.2) which, in the notation of [9], may be written as

$$e^{-u_e/2} = \frac{|z|^2 + (\lambda/8)N}{K}$$

where

$$N = \sum_{n=0}^{n-1} \lambda^i N_i(z) + \sum_{i=n}^{4(n-1)} \lambda^i N_i^n(z),$$

$$K = \sum_{i=0}^{n-1} \lambda^i K_i(z) + \sum_{i=n}^{2(n-1)} \lambda^i K_i^n(z)$$

and N_i, N_i^n, K_i, and K_i^n are as given in [10, App. C and D] for the large norm solution. We recall that u_e is a large norm solution of $-\Delta u = \lambda|f'(z)|^2 e^u$ independent of the choice of G_i's where

$$G(z:\lambda) = 1 + \lambda G_1(z) + \cdots + \lambda^{n-1} G_{n-1}(z)$$

provided $G_i'(0) = 0$ for $i = 1, 2, \cdots, n-1$; and that $K_0 \equiv 1$, $K_1 = 2\operatorname{Re} G_1$, $N_0 = |M_0|$, $N_1 = 2\operatorname{Re}(M_0 \overline{M}_1)$, and

$$M_0(z) = -f'(0) + C_0 z + z I_0(z),$$

$$M_1(z) = -2G_1(0) f'(0) + C_1 z + z I_1(z),$$

where

$$I_0(z)=\int_0^z \frac{f'(\hat{z})-f'(0)}{\hat{z}^2}\,d\hat{z},$$

$$I_1(z)=\int_0^z \frac{2G_1(\hat{z})f'(\hat{z})-2G_1(0)f'(0)}{\hat{z}^2}\,d\hat{z}.$$

It is easy to see that u with $n=1$ ($H=0$, $G=1$) is a first order asymptotic solution (this will be made precise later) of the PDE

$$-\Delta u=\lambda|f'(z)|^2(e^u+\sigma(x)e^{-u}) \tag{2.2}$$

as well as the BC

$$u=0 \quad \text{on } |z|=1. \tag{2.3}$$

Higher order solutions will be obtained by choosing pairs H_i, G_i successively; the H_i's to solve (2.2) asymptotically, and the G_i's to solve (2.3) asymptotically.

More precisely, by substituting (2.1) into (2.2) we obtain:

$$-\Delta u_e-\Delta H=\lambda|f'|^2(\exp(u_e+H)+\sigma(x)\exp(-u_e-H))$$

which may be rewritten as:

$$-\Delta H=\lambda|f'|^2(e^{u_e}(e^H-1)+\sigma e^{-u_e}e^{-H}).$$

This can be shown to be equivalent to $E(z:\lambda)=0$ where

$$E(z:\lambda)=(\Delta H)\left(|z|^2+\frac{\lambda}{8}N\right)^2K^2+\lambda|f'|^2\left(K^4(e^H-1)+\sigma\left(|z|^2+\frac{\lambda}{8}N\right)^4e^{-H}\right). \tag{2.4}$$

If $\|E\|=\max_{|z|\le 1}|E(z:\lambda)|=O(\lambda^n)$ and $\max_{|z|=1}|u(x:\lambda)|=O(\lambda^n)$ we will say that $u(x:\lambda)$ is an asymptotic solution of (P4) of order n. We will see that a solution of order 3 is sufficient. Hence we need only compute H_1, G_1 and H_2, G_2.

Expanding the terms of (2.4) we obtain:

$$\left(|z|^2+\frac{\lambda}{8}N\right)^2=|z|^4+\frac{\lambda}{4}|z|^2|M_0|^2+\frac{\lambda^2}{64}|M_0|^4+\frac{\lambda^2}{2}|z|^2\operatorname{Re}(M_0\overline{M}_1)+O(\lambda^3),$$

$$\left(|z|^2+\frac{\lambda}{8}N\right)^4=|z|^8+\frac{\lambda}{2}|z|^6|M_0|^2+O(\lambda^2),$$

$$K^2=1+4\lambda\operatorname{Re}G_1+4\lambda^2(\operatorname{Re}G_1)^2+4\lambda^2\operatorname{Re}G_2+2\lambda^2|G_1|^2+O(\lambda^3),$$

$$K^4=1+8\operatorname{Re}G_1+O(\lambda^2),$$

$$e^H=1+\lambda H_1+\lambda^2H_2+\tfrac{1}{2}\lambda^2H_1^2+O(\lambda^3),$$

$$e^{-H}=1-\lambda H_1+O(\lambda^2).$$

We can obtain $\|E\|=O(\lambda^3)$ by requiring

$$(\Delta H_1)|z|^4+\sigma|f'|^2|z|^8=0,$$

$$(\Delta H_2)|z|^4+(\Delta H_1)\left(\tfrac{1}{4}|z|^2|M_0|^2+4|z|^4\operatorname{Re}G_1\right)+|f'|^2H_1$$
$$+\sigma|f'|^2\left(\tfrac{1}{2}|z|^6|M_0|^2-H_1|z|^8\right)=0$$

which may be rewritten as

$$-\Delta H_i=\mathcal{Q}_i(z), \qquad i=1,2, \tag{2.5}$$

where, after simplifying

$$\mathcal{Q}_1=\sigma|f'|^2|z|^4,$$

$$\mathcal{Q}_2=|f'|^2\left[\frac{H_1}{|z|^4}+\sigma\left(\frac{1}{4}|z|^2|M_0|^2-|z|^4(H_1+4\operatorname{Re}G_1)\right)\right].$$

By applying the theorem of the Appendix we may choose H_1 of order $|z|^6$ and H_2 of order $|z|^4$ at the origin. For definiteness we select

$$H_i=H_{ip}+H_{iH},\qquad i=1,2,$$

where

$$H_{ip}(z_0)=\iint_{|x|\le 1}g(x_0,x)\mathcal{Q}_i(x)\,dx$$

is the particular solution of (2.5) which has zero boundary values,

$$g(x_0,x)=\frac{1}{4\pi}\ln\left|\frac{z-z_0}{1-\bar{z}_0z}\right|^2$$

is the Green's function for the harmonic ($\lambda=0$) Dirichlet problem associated with (P4), and H_{iH} is chosen by the procedure of the Appendix.

To obtain G_i, $i=1,2$, we apply (2.3) to (2.1) which yields:

$$\left(|z|^2+\frac{\lambda}{8}N\right)^2=K^2e^H\quad\text{for }|z|=1.$$

Equating terms of the same order in the expansions for these terms and simplifying we obtain:

$$2\operatorname{Re}G_i=\tfrac{1}{8}(\mathcal{P}_{i-1}(z)+\mathcal{R}_{i-1}(z)),\qquad i=1,2,$$

where

$$\mathcal{P}_0(z)=|M_0(z)|^2,$$

$$\mathcal{P}_1(z)=2\operatorname{Re}\big(M_0(z)\overline{M_1(z)}\big),$$

$$\mathcal{R}_0(z)=-4H_1(z),$$

$$\mathcal{R}_1(z)=-8|G_1(z)|^2-4H_2(z)+H_1^2(z)-\tfrac{1}{2}H_1(z)|M_0(z)|^2,$$

which may be compared to the results in [10, App. C]. It is easy to see that we can apply the results of [10, App. D] to obtain:

$$G_1(z)=\frac{1}{8}\left\{\frac{|c_0|^2}{2}-\overline{f'(0)}c_0z+\bar{c}_0I_0(z)+\frac{1}{2\pi i}\int_{|z|=1}\mathcal{S}_0(\hat{z})\left(\frac{1}{\hat{z}-z}-\frac{1}{2\hat{z}}\right)d\hat{z}\right\},$$

$$G_2(z)=\frac{1}{8}\left\{c_0\bar{c}_1-\overline{f'(0)}c_1z+\bar{c}_1I_0(z)+\frac{1}{2\pi i}\int_{|z|=1}\mathcal{S}_1(z)\left(\frac{1}{\hat{z}-z}-\frac{1}{2\hat{z}}\right)d\hat{z}\right\}$$

where

$$\mathcal{S}_0(z)=|-f'(0)+zI_0(z)|^2-4H_1(z),$$

$$\mathcal{S}_1(z)=2\operatorname{Re}[M_0(z)(-2f'(0)G_1(0)+I_1(z))]-8|G_1(z)|^2$$
$$-4H_2(z)+H_1^2(z)-\tfrac{1}{2}H_1(z)|M_0(z)|^2.$$

As in [10, App. D], to obtain $G_i'(0)=0$, $i-1,2$, we require

$$c_{i-1}=\frac{f'(0)Z_{i-1}+(f'''(0)/2)\overline{Z}_{i-1}}{|f'(0)|^2-\frac{1}{4}|f'''(0)|^2}$$

where

$$Z_{i-1}=\frac{1}{2\pi i}\int_{|\hat{z}|=1}\mathbb{S}_{i-1}(\hat{z})\,\frac{d\hat{z}}{\hat{z}^2},\qquad i=1,2.$$

Summarizing our results, we have

THEOREM 3. *A large norm asymptotic solution of* (P4) *can be obtained in the form* (2.1) *to order* $n=3$ *by choosing successively* H_1, G_1, H_2 *and* G_2 *as indicated above. That is*:

$$\max_{|z|\le 1}|E(z;\lambda)|=O(\lambda^3),\qquad \max_{|z|=1}|u(z:\lambda)|=O(\lambda^3)$$

where $E(z;\lambda)$ *is given by* (2.4) *and*

$$\max_{|z|\le 1}|u(z:\lambda)|=0\left(\ln\frac{1}{\lambda}\right).$$

3. A modified Newton's method for an equivalent problem. In this section we will show that the large norm asymptotic solution for (P4) given in the previous section has the ability to generate an exact solution via a modified Newton's iteration scheme.

We convert (P4) to the equivalent integral equation ($u\in C(U)$ with $\|u\|=\max_{|x|\le 1}|u(x)|$):

$$u=\mathbb{K}u,$$

where

$$\mathbb{K}=\mathbb{K}_1+\mathbb{K}_2,$$

$$(\mathbb{K}_1u)(x_0)=\int_{|x|\le 1}g(x_0,x)\lambda e^{u(x)}|f'(z)|^2\,dx,$$

$$(\mathbb{K}_2u)(x_0)=\int_{|x|\le 1}g(x_0,x)\lambda\sigma(x)e^{-u(x)}|f'(z)|^2\,dx.$$

As in [13] and [10] we use a modified Newton's method of the form

$$u_{n+1}=\mathbb{S}(u_n) \tag{3.1}$$

where

$$\mathbb{S}u=(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}(\mathbb{K}u-\mathbb{K}'_{u_0}u)$$

and $\mathbb{K}'_{u_0}$ is the Fréchet derivative of $\mathbb{K}$ evaluated at u_0. However, now

$$\mathbb{K}'_{u_0}=\mathbb{K}'_{1u_0}+\mathbb{K}'_{2u_0} \tag{3.2}$$

where

$$(\mathbb{K}_{1u_0}h)(x_0)=\int_{|x|\le 1}g(x_0,x)\lambda e^{u_0(x)}h(x)|f'(z)|^2\,dx,$$

$$(\mathbb{K}_{2u_0}h)(x_0)=\int_{|x|\le 1}g(x_0,x)\lambda\sigma(x)e^{-u_0(x)}h(x)|f'(z)|^2\,dx.$$

Now

$$\mathbb{S}'_u h=(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}(\mathbb{K}_u h-\mathbb{K}'_{u_0}h),$$

and since

$$\mathbb{K}'_u h-\mathbb{K}'_{u_0}h=\mathbb{K}'_{1u_0}[(e^{u-u_0}-1)h]+\mathbb{K}'_{2u_0}[(e^{-(u-u_0)}-1)h]$$

we have that

$$\|\mathbb{S}'_u\|\le\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\mathbb{K}'_{1u_0}\|\,\|e^{u-u_0}-1\|+\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\mathbb{K}'_{2u_0}\|\,\|e^{-(u-u_0)}-1\|.$$

Now since

$$\|e^{u-u_0}-1\|\le e^t-1 \quad \text{if } \|u-u_0\|\le t,$$
$$\|e^{-(u-u_0)}-1\|\le e^t-1 \quad \text{if } \|u-u_0\|\le t,$$

we have

$$\|\mathbb{S}'_u\|\le\Gamma(e^t-1) \quad \text{if } \|u-u_0\|\le t$$

where Γ satisfies

$$\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\mathbb{K}'_{1u_0}\|+\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\mathbb{K}'_{2u_0}\|\le\Gamma. \tag{3.3}$$

Now since

$$\mathbb{S}(u)=u-(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}(u-\mathbb{K}u),$$
$$\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\|\le 1+\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\mathbb{K}'_{u_0}\|,$$

we also have

$$\|\mathbb{S}(u_0)-u_0\|\le(1+\Gamma)\|u_0-\mathbb{K}_{u_0}\|.$$

Hence,

$$\phi(t)=\Gamma(e^t-t-1)+\phi(0)$$

majorizes $\mathbb{S}$ [12, p. 260] provided that $\phi(0)$ satisfies

$$(1+\Gamma)\|u_0-\mathbb{K}u_0\|\le\phi(0).$$

It is easy to show that $\phi(t)=t$ has a unique positive solution $t^*\le\ln[1+(1/\Gamma)]$ if

$$\|u_0-\mathbb{K}u_0\|\le\ln\left[1+\frac{1}{\Gamma}\right]-\frac{1}{1+\Gamma}.$$

Hence applying the result of Kantorovitch [12, p. 260] we have (compare with [10, Thm. 4]) the following:

THEOREM 4. *If* $\|u_0-Ku_0\|\le\ln[1+(1/\Gamma)]-1/(1+\Gamma)$ *where* Γ *satisfies* (3.3), *then the modified Newton's method* (3.1) *will converge to a solution* u^* *of* (P4) *such that*

$$\|u_0-u^*\|\le\ln\left[1+\frac{1}{\Gamma}\right].$$

We will always use the asymptotic solution (2.1) as u_0 in (3.1). To apply Theorem 4 we require estimates of $\|u_0-\mathbb{K}u_0\|$ and Γ.

4. Estimate for $\|u_0-\mathbb{K}u_0\|$. Applying the representation formula

$$u=-\int_{|x|\le 1} g(\cdot,x)\Delta u\,dx-\int_{|x|=1} u\frac{\partial g}{\partial n}\,ds$$

to $u_0=u_e+H$ we obtain

$$u_0=-\int_{|x|\le 1} g(x_0,x)\left(-\lambda|f'|^2e^{u_e}+\Delta H\right)dx-\int_{|x|=1} u_0\frac{\partial g}{\partial n}\,ds.$$

Hence

$$u_0-\mathbb{K}u_0=\int_{|x|\le 1} g(x_0,x)\left\{\lambda|f'|^2\left[e^{u_e}(1-e^H)-\sigma e^{-u_e}e^{-H}\right]-\Delta H\right\}dx$$
$$-\int_{|x|=1} u_0\frac{\partial g}{\partial n}\,ds,$$

which may be rewritten using the definition of u_e and $E(z:\lambda)$ as:

$$u_0-\mathbb{K}u_0=\int_{|x|=1} u_0\left(-\frac{\partial g}{\partial n}\right)ds+\int_{|x|\le 1} g(x_0,x)\frac{E(z:\lambda)\,dx}{K^2\left(|z|^2+\frac{\lambda}{8}N\right)^2}.$$

From this we can obtain:

$$|u_0-\mathbb{K}u_0|\le\max_{|x|=1}|u_0(x)|+\left\|\frac{E(z:\lambda)}{K^2}\right\|\int_{|x|\le 1}\frac{g(x_0,x)\,dx}{\left(|z|^2+\frac{\lambda}{8}N\right)^2}.$$

Using the techniques of [10, App. E] it can be shown that

$$\int_{|x|\le 1}\frac{g(x_0,x)\,dx}{\left(|z|^2+\frac{\lambda}{8}N\right)^2}\le M_1\int_{|x|\le 1}\frac{g(x_0,x)\,dx}{\left(|z|^2+\frac{\lambda}{8}|f'(0)|^2\right)^2}$$
$$=O\left(\lambda^{-1/2}\ln\frac{1}{\lambda}\right).$$

Now since we are taking u_0 to be a third order solution and since $K\to 1$ as $\lambda\to 0$, we have

$$\|u_0-\mathbb{K}u_0\|=O\left(\lambda^{5/2}\ln\frac{1}{\lambda}\right). \tag{4.1}$$

5. Estimate for Γ. We estimate $\|\mathbb{K}'_{1u_0}\|$, $\|\mathbb{K}'_{2u_0}\|$ and $\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\|$. Using the definition of u_e, we have that

$$e^{u_0}=e^{u_e+H}=e^{u_e}+\lambda e^{u_e}H_1+e^{u_e}\mathfrak{M}_{R_1},$$
$$e^{-u_0}=e^{-u_e-H}=|z|^4+\mathfrak{M}_{R_2},$$

where

$$\|\mathfrak{M}_{R_1}\|=O(\lambda^2),\qquad \|\mathfrak{M}_{R_2}\|=O(\lambda).$$

Hence we obtain

$$\mathbb{K}'_{1u_0}=\mathbb{K}_e+\mathbb{K}_{\eta_3}+\mathbb{K}_{R_1},$$
$$\mathbb{K}'_{2u_0}=\mathbb{K}_{\eta_4}+\mathbb{K}_{R_2},$$

where $\mathbb{K}_e$, $\mathbb{K}_{\eta_3}$, $\mathbb{K}_{R_1}$, $\mathbb{K}_{\eta_4}$, and $\mathbb{K}_{R_2}$ are linear integral operators with kernels given respectively by:

$$g(x_0,x)\lambda e^{u_e}|f'(z)|^2,$$
$$g(x_0,x)\lambda^2 e^{u_e}H_1(z)|f'(z)|^2,$$
$$g(x_0,x)\lambda e^{u_e}\mathfrak{M}_{R_1}(x)|f'(z)|^2,$$
$$g(x_0,x)\lambda|z|^4\sigma|f'(z)|^2,$$
$$g(x_0,x)\lambda\mathfrak{M}_{R_2}(x)\sigma|f'(z)|^2.$$

We note that $\mathbb{K}_e$ is the $\mathbb{K}'_{u_0}$ of [10], and hence we easily obtain:

$$\|\mathbb{K}_e\| = O\left(\ln\frac{1}{\lambda}\right), \qquad \|\mathbb{K}_{\eta_4}\| = O(\lambda), \tag{5.1}$$
$$\|\mathbb{K}_{\eta_3}\| = O\left(\lambda\ln\frac{1}{\lambda}\right), \qquad \|\mathbb{K}_{R_2}\| = O(\lambda^2).$$
$$\|\mathbb{K}_{R_1}\| = O\left(\lambda^2\ln\frac{1}{\lambda}\right),$$

Thus

$$\|\mathbb{K}'_{1u_0}\| = O\left(\ln\frac{1}{\lambda}\right), \qquad \|\mathbb{K}'_{2u_0}\| = O(\lambda). \tag{5.2}$$

Furthermore

$$\mathbb{K}'_{u_0} = \mathbb{K}_e + \mathbb{K}_{\eta_3} + \mathbb{K}_{\eta_4} + \mathbb{K}_{R_1} + \mathbb{K}_{R_2}.$$

Hence we may apply the results of [10, App. E] where we need only redefine $\mathbb{L}$ as

$$\mathbb{L} = \mathbb{K}_1 + \mathbb{K}_{\eta_1} + \mathbb{K}_{\eta_2} + \mathbb{K}_{\eta_3} + \mathbb{K}_{\eta_4} + \mathbb{K}_R + \mathbb{K}_{R_1} + \mathbb{K}_{R_2}.$$

where $\mathbb{K}_1$, $\mathbb{K}_{\eta_1}$, $\mathbb{K}_{\eta_2}$, and $\mathbb{K}_R$ are as given in [10]. Note in particular that we still have

$$\|\mathbb{L}\| = O\left(\lambda\left(\ln\frac{1}{\lambda}\right)^2\right).$$

Hence we wish an asymptotic solution to

$$\sum_{j=2}^{4} a_{ij}c_j = -w_j$$

where

$$w_j = (\mathbb{I}-\mathbb{M})^{-1}\left((\mathbb{I}-\mathbb{N})^{-1}w, \phi_j\right),$$
$$a_{ij} = (\mathbb{L}\phi_j, \phi_i) + (\mathbb{L}\mathbb{M}\phi_j, \phi_i) + O\left(\lambda^{3/2}\left(\ln\frac{1}{\lambda}\right)^6\right) \quad \text{for } i \text{ or } j \text{ in } \{2,3\},$$
$$a_{44} = \frac{3}{2}(\ln u) + O\left(\left(\ln\frac{1}{\lambda}\right)^{-1}\right),$$
$$\mathbb{N} = \mathbb{K}_0 - \sum_{j=2}^{4} \Lambda_j\phi_j(\cdot, \phi_j),$$
$$\mathbb{M} = (\mathbb{I}-\mathbb{N})^{-1}\mathbb{L},$$

and $\mathbb{K}_0$ and Λ_j are as in [10, App. E]. However, using (5.1) we now have

$$(\mathbb{L}\phi_j,\phi_i)_\rho=(\mathbb{K}\phi_j,\phi_i)+\Lambda_i(\eta_1\phi_j,\phi_i)_\rho+(\eta_1\phi_j,\mathbb{K}_1\phi_i)+\Lambda_i(\eta_2\phi_j,\phi_i)+(\mathbb{K}_{\eta_3}\phi_j,\phi_i)$$
$$+(\mathbb{K}_{\eta_4}\phi_j,\phi_i)+O\left(\lambda^{3/2}\ln\frac{1}{\lambda}\right).$$

Now since

$$\mathbb{K}_{\eta_3}\phi_j=(\mathbb{K}_0+\mathbb{K}_1)[\lambda H_1\phi_j],$$

we have

$$(\mathbb{K}_{\eta_3}\phi_j,\phi_i)=\Lambda_i(\lambda H_1\phi_j,\phi_i)+O(\lambda^{3/2}).$$

We recall that $H_1(z)$ is of order $|z|^6$ at the origin. Hence it can be shown by the methods of [10, App. E] that

$$(\lambda\sigma H_1\phi_j,\phi_i)=O(\lambda^2).$$

Similarly

$$(\mathbb{K}_{\eta_4}\phi_j,\phi_i)=(\lambda|z|^4\sigma\phi_j,\phi_i)=O(\lambda^2).$$

Hence we see that asymptotically $(\mathbb{L}\phi_j,\phi_i)$ is as in [10, App. E]. Using (5.1) it is easy to see that asymptotically $(\mathbb{L}\mathbb{M}\sigma_j,\phi_i)$ is as in [10, App. E]. Hence we obtain

$$\|c_i\|\leq\text{constant}\lambda^{-1}\|w\|,$$

and hence

$$\|(\mathbb{I}-\mathbb{K}'_{u_0})^{-1}\|=O\left(\frac{1}{\lambda}\right).$$

Thus

$$\Gamma=O\left(\frac{\ln(\frac{1}{\lambda})}{\lambda}\right).$$

6. Convergence of the modified Newton's method. We recall that for $\Gamma>1$ we have

$$\ln\left(1+\frac{1}{\Gamma}\right)-\frac{1}{1+\Gamma}=\frac{1}{2}\left(\frac{1}{\Gamma}\right)^2-\frac{2}{3}\left(\frac{1}{\Gamma}\right)^3+\cdots.$$

Now recalling (4.1) and noting that

$$\lambda^{5/2}\ln\frac{1}{\lambda}=o\left(\lambda^2\ln\left(\frac{1}{\lambda}\right)^{-2}\right),$$

we apply Theorem 4 to obtain

THEOREM 5. *If the large norm asymptotic solution* (2.1) *with* $n=3$ *is used as a first approximation in the modified Newton's method* (3.1), *then* u_n *will converge to a unique large norm solution* u^* *such that*

$$\|u_0-u^*\|=O(\lambda).$$

7. A third solution of (P1) if ψ is a negative constant. Suppose that ψ is a negative constant and that λ is sufficiently small so that we may obtain a large norm "positive" solution to both (P1) and

$$\text{(P5)} \qquad -\Delta u=\lambda(-\psi)\left(e^{u}+\left(\frac{1}{\psi}\right)e^{-u}\right),$$
$$u=-\phi,$$

by the technique of the previous sections. We call such solutions "positive" since at the point $f_{\Omega}(\delta)$ the solution increases positively without bound as $\lambda\to 0$. If $\hat{u}$ is the solution of (P5) obtained, then $-\hat{u}$ is a large norm "negative" solution of (P1).

Appendix. The order of the zero at the origin for Poisson's equation. In this appendix we consider the Poisson equation

$$\text{(A.1)} \qquad \Delta u=g(x), \qquad x\in\Omega,$$

where $x=(x_1,x_2)$, Ω is a bounded simply connected region containing the origin, and g is real analytic in Ω. We prove the following theorem on the order of the zero at the origin.

THEOREM A.1. *If $g(z)=O(|x|^{2k})$ with $k\geq 0$, then there exists an analytic solution of* (A.1) *such that $u=O(|x|^{2k+2})$.*

Proof. Since (A.1) is elliptic, there exists an analytic solution [8] in Ω. Denote this solution by

$$u_p(x)=\sum_{|\alpha|\leq 2k+1} a_\alpha x^\alpha+O\left(|x|^{2k+2}\right)$$

where we employ the standard notation and

$$a_\alpha=\left.\frac{D^\alpha u_p}{\alpha!}\right|_{x=[0,0]}.$$

Letting $x_1=r\cos\theta$, $x_2=r\sin\theta$,

$$\text{(A.2)} \qquad H(r,\theta)=\frac{A_0}{2}+\sum_{l=1}^{2k+1} r^l\left(A_l\cos l\theta+B_l\sin l\theta\right),$$

$$\text{(A.3)} \qquad \Theta_l(\theta)=\sum_{i=1}^{l-1}\left(C_i^l\cos i\theta+D_i^l\sin i\theta\right),$$

we may rewrite u_p as

$$u_p=H(r,\theta)+\sum_{l=2}^{2k+1} r^l\Theta_l+O(r^{2k+2})$$

where the C_i^l's, D_i^l's, and B_l's are defined in terms of the a_α's. Hence

$$\Delta u_p=\Delta\left(\sum_{l=2}^{2k+1} r^l\Theta_l\right)+O(r^{2k}),$$

so that substituting into (A.1) we have

$$g(x)=\sum_{l=2}^{2k+1} r^{l-2}\left(l^2\Theta+\frac{d^2\Theta_l}{d\theta^2}\right)+O(r^{2k}).$$

Since $g=O(r^{2k})$, we must have

$$\frac{d^2\Theta_l}{d\theta^2}+l^2\Theta=0. \tag{A.4}$$

But there are no solutions of (A.4) of the form (A.3) except $C_i^l=D_i^l=0$ for $0\le i\le l-1$. Hence by subtracting the harmonic function (A.2) from u_p we obtain

$$u=u_p-H=O(r^{2k+2})$$

which is a solution of (A.1) of order $|x|^{2k+2}$.

REFERENCES

[1] R. ARIS, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Clarendon Press, Oxford, 1975.

[2] C. BANDLE, *Existence theorems, qualitative results and a priori bounds for a class of nonlinear Dirichlet problems*, Arch. Rational Mech. Anal., 58 (1975), pp. 219–238.

[3] _____, *Isoperimetric inequalities for a nonlinear eigenvalue problem*, Proc. Amer. Math. Soc., 56 (1976), pp. 243–246.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Interscience, New York, 1962.

[5] M. G. CRANDALL AND P. N. RABINOWITZ, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Rational Mech. Anal., 58 (1975), pp. 207–218.

[6] G. M. GOLUZIN, *Geometric Theory of Functions of a Complex Variable*, American Mathematical Society, Providence, RI, 1969.

[7] J. B. KELLER AND D. S. COHEN, *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), pp. 1361–1376.

[8] C. B. MORREY, *On the analyticity of the solutions of analytic nonlinear elliptic systems of partial differential equations*, I *and* II, Amer. J. Math., 80 (1958), pp. 198–237.

[9] J. L. MOSELEY, *Asymptotic solutions for a Dirichlet problem with an exponential nonlinearity*, this Journal, 14 (1983), pp. 719–735.

[10] _____, *On asymptotic solutions for a Dirichlet problem with an exponential nonlinearity*, Rep. AMR1, West Virginia Univ., Morgantown, 1981.

[11] N. L. SCHRYER, *Solution of monotone nonlinear elliptic boundary value problems*, Numer. Math., 18 (1972), pp. 336–344.

[12] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.

[13] V. H. WESTON, *On the asymptotic solution of a partial differential equation with an exponential nonlinearity*, this Journal, 9 (1978), pp. 1030–1053.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0009 $01.25/0

CONSTANT-RATE HARVESTING OF AGE-STRUCTURED POPULATIONS*

FRED BRAUER†

Abstract. We consider the nonlinear age-dependent population growth model introduced by Gurtin and MacCamy [Arch. Rat. Mech. Anal., 54 (1974), pp. 281–300] with birth and death moduli depending on age and total population size, to which is added a harvest of members at a rate which is constant in time but may depend on the age of members being harvested. We formulate the problem as a pair of equations for the total population size and the birth rate, and discuss the behaviour of solutions when the birth and death moduli depend on either the age or the total population size, but not both.

1. The study of age-dependent population models goes back to the work of Sharpe and Lotka [27] in which the birth rate is expressed as the solution of a linear integral equation—the renewal equation. Introduction of the age density function leads to a partial differential equation usually known as the von Foerster equation [33], although it can be traced back to the work of MacKendrick [19].

A nonlinear variant of the MacKendrick equation was proposed by Gurtin and MacCamy [11] with the birth and death moduli depending on the age and also on the total population size. This model was then transformed into a pair of functional equations for the birth rate and the total population size. Various other generalizations have been studied as well. For example, birth and death moduli which depend on the age density function have been considered by Griffel [10] and Sinestrari [28], birth moduli which depend on the birth rate have been considered by Rorres [21], [22], [23] and Swick [29], [30], [31], [32], and models which incorporate response delays in the birth and death rates have been examined by Cushing [5], [6]. While any of these types of models may be appropriate to a specific population, we shall deal only with the Gurtin–MacCamy model. Many of the techniques can be modified and applied to other types of models.

We shall incorporate a harvest of members with a preassigned age structure and constant total time rate into the basic model. Constant effort harvesting with an effort which may depend on the age and the total population size has been studied by Sánchez [25], [26]. Rorres and Fair [24], Getz [8], [9] and Gurtin and Murphy [14], [15], including questions of optimization. Our goal is to study constant-yield harvesting, considering equilibrium age distributions, persistent age distributions and the asymptotic behavior of solutions. We shall establish some results with the aid of various assumptions on the form of the birth and death moduli. If the birth modulus depends only on population size and the death modulus depends only on age, our model reduces to the nonlinear renewal equation and in §4 we describe the known results in this case. For sufficiently small harvest rates there can be an equilibrium age distribution and a condition for stability of an equilibrium can be given. In §5 we consider the case in which both birth and death moduli are functions of age only and show that there cannot be an equilibrium under harvesting; the only possibilities are extinction in finite time and persistent age distributions which may tend to zero or may be unbounded. The case of birth moduli depending on age only and death moduli depending on population size only is considered in §6. The possibilities of extinction in finite time

*Received by the editors February 3, 1982, and in revised form July 15, 1982.
†Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

and unbounded populations also arise in this case. While we cannot rule out bounded population size, we can show that there cannot be an equilibrium under harvesting.

2. Let $\rho(a,t)$ be the density with respect to age of members of a population of age a at time t, so that the number of members with ages between a and $a+\Delta a$ at time t is approximately $\rho(a,t)\Delta a$. Then the total population at time t is

$$P(t)=\int_0^\infty \rho(a,t)\,da. \tag{1}$$

Let $\mu(a,P)$ be the death modulus—the death rate at age a when the total population size is P. In addition, we assume a harvest of members at a rate which is constant in time but may depend on age. We let this rate be $\nu(a)$, so that the total harvest rate is

$$H=\int_0^\infty \nu(a)\,da, \tag{2}$$

which we assume finite. A harvest rate proportional to $\rho(a,t)$ [constant effort harvesting with an effort $E(a,P(t))$ which may depend on the age a and the total population size $P(t)$] may be included in the model by replacing $\mu(a,P)$ by $\mu(a,P)+E(a,P)$, but we shall not consider proportional harvesting further here. Then ρ satisfies

$$\rho_t(a,t)+\rho_a(a,t)+\mu(a,P(t))\rho(a,t)+\nu(a)=0, \tag{3}$$

valid if the partial derivatives $\rho_t(a,t)$ and $\rho_a(a,t)$ exist. The unharvested case $\nu(a)\equiv 0$ is the so-called von Foerster equation [33], although in view of the priority which has been pointed out by Hoppensteadt [16] it is more appropriately described as the MacKendrick equation.

Let $\beta(a,P)$ be the birth-modulus—the average number of offspring per unit time produced by an individual of age a. Then the number of births per unit time at time t is the birth rate $B(t)$, satisfying

$$B(t)=\rho(0,t)=\int_0^\infty \beta(a,P(t))\rho(a,t)\,da, \qquad t>0. \tag{4}$$

To complete the model, we must specify an initial age distribution

$$\rho(a,0)=\phi(a), \qquad 0\le a<\infty, \tag{5}$$

and we require $\int_0^\infty \phi(a)\,da<\infty$ in order to assure a finite initial total population size. We do not, however, insist that (4) and (5) be compatible at $t=0$, $a=0$.

The population growth model now consists of the partial differential equation (3) together with the auxiliary conditions (4), (5). We shall replace the system (3), (4), (5) by a pair of functional equations in the unknown functions $P(t)=\int_0^\infty \rho(a,t)\,da$ (total population size) and $B(t)=\rho(0,t)$ (birth rate). We follow the standard technique for achieving this (see, for example [16] or [11], [12]). We obtain

$$\rho(a,t)=\exp\left(-\int_0^a \mu^*(\alpha)\,d\alpha\right)B(t-a)-\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)d\eta \qquad (t\ge a) \tag{6}$$

and

$$\rho(a,t)=\exp\left(-\int_{a-t}^a \mu^*(\alpha)\,d\alpha\right)\phi(a-t)-\int_{a-t}^a \nu(\eta)\exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)d\eta \qquad (t\le a). \tag{7}$$

Here, $\mu^*(\alpha)$ denotes $\mu(\alpha, P(t-a+\alpha))$.

Substitution of (6) and (7) into (4) now gives

$$B(t)=b(t)+\int_0^t \beta(a,P(t))\exp\left(-\int_0^a \mu^*(\alpha)\,d\alpha\right)B(t-a)\,da-h_1(t), \tag{8}$$

where

$$b(t)=\int_t^\infty \beta(a,P(t))\phi(a-t)\exp\left(-\int_{a-t}^a \mu^*(\alpha)\,d\alpha\right)da \tag{9}$$

and

$$\begin{aligned} h_1(t)&=\int_0^t \beta(a,P(t))\left[\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)d\eta\right]da \\ &\quad+\int_t^\infty \nu(\eta)\left[\int_\eta^{\eta+t} \beta(a,P(t))\exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)da\right]d\eta \\ &=\int_0^\infty \nu(\eta)\left[\int_\eta^{\eta+t} \beta(a,P(t))\exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)da\right]d\eta. \end{aligned} \tag{10}$$

A similar calculation using (1) in place of (4) gives

$$P(t)=p(t)+\int_0^t \exp\left(-\int_0^a \mu^*(\alpha)\,d\alpha\right)B(t-a)\,da-h_2(t) \tag{11}$$

with

$$p(t)=\int_t^\infty \phi(a-t)\exp\left(-\int_{a-t}^a \mu^*(\alpha)\,d\alpha\right)da, \tag{12}$$

$$h_2(t)=\int_0^\infty \nu(\eta)\left[\int_\eta^{\eta+t} \exp\left(-\int_\eta^a \mu^*(\alpha)\,d\alpha\right)da\right]d\eta. \tag{13}$$

From (9), (10), (12), (13) it is easy to see that

$$b(0)=\int_0^\infty \beta(a,P(0))\phi(a)\,da>0,$$

$b(t)\geq 0$ for $0\leq a<\infty$, $b(\infty)=\lim_{t\to\infty} b(t)=0$, that

$$p(0)=\int_0^\infty \phi(a)\,da>0,$$

$p(t)=0$ for $0\leq a<\infty$, $p(\infty)=\lim_{t\to\infty} p(t)=0$, that $h_1(0)=h_2(0)=0$, $h_1(t)\equiv 0$ and $h_2(t)\equiv 0$ for $0\leq t<\infty$ if $\nu(a)\equiv 0$ (no harvest) and $h_1(t)\geq 0$, $h_2(t)\geq 0$ for $0\leq t<\infty$. It should be observed that $b(t)$, $p(t)$, $h_1(t)$ and $h_2(t)$ are functions of the total population size in general, but if μ is a function of a only then $p(t)$ and $h_2(t)$ are independent of P and if μ and β are both functions of a only then $b(t)$ and $h_1(t)$ are independent of P.

The population growth model is now described by the pair of functional equations (8) and (11). A solution of this system yields the age density function $\rho(a,t)$ by substitution in (6) and (7). Thus the system (8), (11) is equivalent to the original problem (3), (4), (5), and this is our first main result, analogous to the corresponding result in [11].

THEOREM 1. *The pair of functional equations* (8), (11) *for the birth rate* $B(t)$ *and the total population size* $P(t)$ *is equivalent to the problem* (3), (4), (5).

If the death modulus is independent of age, and is a function of the total population size only, $\mu=\mu(P)$, then the equation (11) may be replaced by an ordinary differential equation, much as in the unharvested case, namely

$$P'=B-P\mu(P)-H, \tag{14}$$

where $H=\int_0^\infty \nu(a)\,da$ is the total harvest rate. The system (8), (11) may be replaced by the system (8), (14) in this case, a fact first observed by Gurtin and MacCamy [12]. It is also possible to use (14) to eliminate B from (8) and describe the model by a single equation for P, a technique which has been exploited by Cushing [5], [6].

The equation (14) may also be derived directly from (11) by differentiation if $\mu=\mu(P)$.

A different major simplification can be made if the birth modulus is independent of age and is a function of the total population size P only, $\beta=\beta(P)$. In this case, (4) becomes

$$\begin{aligned} B(t)&=\int_0^\infty \beta(P(t))\rho(a,t)\,da \\ &=\beta(P(t))\int_0^\infty \rho(a,t)\,da=P(t)\beta(P(t)) \end{aligned} \tag{15}$$

Then (11) may be written

$$P(t)=p(t)+\int_0^t \exp\left(-\int_0^a \mu^*(\alpha)\,d\alpha\right)P(t-a)\beta(P(t-a))\,da-h_2(t)$$

or

$$P(t)=p(t)+\int_0^t \exp\left(-\int_0^a \mu^*(\alpha)\,d\alpha\right)g(P(t-a))\,da-h_2(t), \tag{16}$$

where $g(P)=P\beta(P)$. The system (8), (11) is then equivalent to the single equation (16) for P, together with the formula (15) for B in terms of P.

In general, the system (8), (11), or even the system (8), (14) in the case $\mu=\mu(P)$, cannot be analyzed completely. However, it is possible to obtain useful information about the behavior of solutions in some special cases. As the concept of an equilibrium age distribution is of use in some of these special cases, we shall discuss it briefly and then proceed to examine some of the cases in which useful information can be obtained.

3. An equilibrium age distribution for the problem (3), (4), (5) is defined to be a solution $\rho(a)$ which is independent of t. It is clear from (1) and (4) that the population size and birth rate corresponding to an equilibrium age distribution are constant. Conversely, it is clear from (6) that the age distribution corresponding to a constant solution of the system (8), (11) is an equilibrium age distribution.

In view of the equivalence between equilibrium age distributions and constant birth rates and population sizes, it is natural to inquire what values (B_0, P_0) are possible constant solutions of (10), (13). The corresponding age distribution satisfies the ordinary differential equation

$$\rho'(a)=-\mu(a,P_0)\rho(a)-\nu(a), \qquad \rho(0)=B_0,$$

By solving this equation and using (1) and (4) we obtain the pair of conditions

(17)
$$B_0\left[\int_0^\infty \beta(a,P_0)\exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da-1\right]$$
$$=\int_0^\infty \beta(a,P_0)\left[\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu(\alpha,P_0)\,d\alpha\right)d\eta\right]da,$$

(18)
$$P_0=B_0\int_0^\infty \exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da-\int_0^\infty\left[\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu(\alpha,P_0)\,d\alpha\right)d\eta\right]da.$$

The conditions (17) and (18) may also be obtained by assuming a constant solution (B_0,P_0) of (8), (11) and letting $t\to\infty$. For this reason we write

$$h_1(\infty)=\int_0^\infty \beta(a,P_0)\left[\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu(\alpha,P_0)\,d\alpha\right)d\eta\right]da,$$
$$h_2(\infty)=\int_0^\infty\left[\int_0^a \nu(\eta)\exp\left(-\int_\eta^a \mu(\alpha,P_0)\,d\alpha\right)d\eta\right]da,$$

as $h_1(\infty)=\lim_{t\to\infty}h_1(t)$, $h_2(\infty)=\lim_{t\to\infty}h_2(t)$ if P is constant; note that $h_1(\infty)$ and $h_2(\infty)$ depend on P_0. With this notation, the conditions (17), (18) become

(19)
$$B_0\left[\int_0^\infty \beta(a,P_0)\exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da-1\right]=h_1(\infty),$$

(20)
$$P_0=B_0\int_0^\infty \exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da-h_2(\infty).$$

If there is no harvest, $h_1(\infty)=h_2(\infty)=0$. In this case (19) becomes

$$\int_0^\infty \beta(a,P_0)\exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da=1,$$

stating that the average number of offspring per member when the population size is P_0 must be 1, and this is an equation which may be solved for P_0 (see [11],[12]). Then (20) expresses B_0 in terms of P_0. If there is a harvest, $h_1(\infty)>0$, and this implies

$$\int_0^\infty \beta(a,P_0)\exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da>1.$$

In this case, (19) and (20) cannot be uncoupled so readily, although it is possible to eliminate B_0 and obtain the equation

$$h_1(\infty)\int_0^\infty \exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da$$
$$=[P_0+h_2(\infty)]\left[\int_0^\infty \beta(a,P_0)\exp\left(-\int_0^a \mu(\alpha,P_0)\,d\alpha\right)da-1\right]$$

for P_0.

The relation (19) may also be derived from the equation

$$\int_0^\infty \mu(\alpha, P_0)\rho(a)\,da+\int_0^\infty \nu(a)\,da=\int_0^\infty \beta(a, P_0)\rho(a)\,da=B_0,$$

which expresses the fact that at equilibrium the birth rate must be equal to the death rate plus the harvest rate.

As we have seen, if μ is a function of population size only, $\mu=\mu(P)$, then (11) may be replaced by the ordinary differential equation (14). In this case, a constant solution (B_0, P_0) must satisfy

$$B_0=P_0\mu(P_0)+\int_0^\infty \nu(a)\,da,$$

a condition which may be used in place of (20) (and indeed one to which (20) may be reduced).

4. If the birth modulus is a function of population size only, $\beta=\beta(P)$, and the death modulus is a function of age only, $\mu=\mu(a)$, then the system (10), (13) can be reduced to a single Volterra equation, the nonlinear renewal equation

$$P(t)=p(t)+\int_0^t \pi_0(a)g(P(t-a))\,da-h_2(t), \tag{21}$$

where

$$\pi_0(a)=\exp\left(-\int_0^a \mu(\alpha)\,d\alpha\right), \qquad g(P)=P\beta(P), \tag{22}$$

$$\begin{aligned} p(t)&=\int_t^\infty \phi(a-t)\exp\left(-\int_{a-t}^a \mu(\alpha)\,d\alpha\right)da \\ &=\int_t^\infty \frac{\phi(a-t)\pi_0(a)}{\pi_0(a-t)}\,da, \end{aligned}$$

$$\begin{aligned} h_2(t)&=\int_0^\infty \nu(\eta)\left[\int_\eta^{\eta+t}\exp\left(-\int_\eta^a \mu(\alpha)\,d\alpha\right)da\right]d\eta \\ &=\int_0^\infty \frac{\nu(\eta)}{\pi_0(\eta)}\left[\int_\eta^{\eta+t}\pi_0(a)\,da\right]d\eta. \end{aligned} \tag{23}$$

Population models of this type were introduced by Cooke and Yorke [4] and have also been studied by the author [1], [2], [3] although without making the observation that the age distribution may be derived from a solution of the nonlinear renewal equation via (17) and (8).

For (21) it is known that if

$$\left[\int_0^\infty \pi_0(a)\,da\right]\left[\limsup_{P\to\infty}\beta(P)\right]<1,$$

then every nonnegative solution is bounded on $0\le t<\infty$ [1]. Further if $g'(P)\int_0^\infty \pi_0(a)\,da$ is not identically equal to 1 on any P-interval, then every nonnegative solution tends to a limit P_∞ as $t\to\infty$, where

$$P_\infty=-h_2(\infty)+g(P_\infty)\int_0^\infty \pi_0(a)\,da,$$

with

$$h_2(\infty)=\lim_{t\to\infty} h_2(t)=\int_0^\infty \frac{\nu(\eta)}{\pi_0(\eta)}\left[\int_\eta^\infty \pi_0(a)\,da\right]d\eta,$$
$$=\int_0^\infty \pi_0(a)\left[\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right]da$$

[17], [18]. If $h_2(\infty)$ is sufficiently large that the curve $y=g(P)$ and the line $y=(P+h_2(\infty))/\int_0^\infty \pi_0(a)\,da$ in the P-y plane do not intersect, then there are no possible values of P_∞, and the solution $P(t)$ of (21) must reach zero in finite time and then become negative. We interpret a negative population size as indicating extinction of the population, and once a population size has become negative we do not pursue it further. Stability of an equilibrium P_∞ in the sense of relative insensitivity to perturbations requires [3]

$$g'(P_\infty)\int_0^\infty \pi_0(a)\,da<1.$$

Since here $g(P)=P\beta(P)$, the equilibrium condition may be written

$$P_\infty=h_2(\infty)+P_\infty\beta(P_\infty)\int_0^\infty \pi_0(a)\,da$$

and the stability condition may be written

$$[P_\infty\beta'(P_\infty)+\beta(P_\infty)]\int_0^\infty \pi_0(a)\,da<1,$$
$$P_\infty\beta'(P_\infty)\int_0^\infty \pi_0(a)\,da=1-\left[1+\frac{h_2(\infty)}{P_\infty}\right]=-\frac{h_2(\infty)}{P_\infty},$$

or

$$P_\infty^2\beta'(P_\infty)\int_0^\infty \pi_0(a)\,da+h_2(\infty)<0.$$

In particular, $\beta'(P_\infty)<0$ is necessary for stability, and if there is no harvest so that $h_2(\infty)=0$, $\beta'(P_\infty)<0$ is necessary and sufficient for stability.

5. The classical situation originally studied by MacKendrick [19] assumed that both the birth and death moduli were functions of age only, $\beta=\beta(a)$, $\mu=\mu(a)$. In this case, the system (8), (11) becomes

$$B(t)=b(t)+\int_0^t \beta(a)\pi_0(a)B(t-a)\,da-h_1(t), \tag{24}$$
$$P(t)=p(t)+\int_0^t \pi_0(a)B(t-a)\,da-h_2(t),$$

where

$$\pi_0(a)=\exp\left(-\int_0^a \mu(\alpha)\,d\alpha\right), \tag{25}$$

$p(t)$ and $h_2(t)$ are given by (25) and (26) respectively, and

$$b(t)=\int_t^\infty \frac{\beta(a)\pi_0(a)\phi(a-t)}{\pi_0(a-t)}da,$$

$$h_1(t)=\int_0^\infty \frac{\nu(\eta)}{\pi_0(\eta)}\left[\int_n^{n+t}\beta(a)\pi_0(a)\,da\right]d\eta.$$

We may regard (24) as a single linear Volterra integral equation for $B(t)$ together with an explicit formula for $P(t)$ in terms of $B(t)$. In terms of $B(t)$, the age distribution $\rho(a,t)$ is given by the special case of (6)

$$\rho(a,t)=\pi_0(a)\left[B(t-a)-\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right], \tag{26}$$

for $t\geq a$. The case of no harvesting, $h_1(t)\equiv 0$, $h_2(t)\equiv 0$ has been studied by Gurtin and MacCamy [12], who have established the existence of persistent age distributions of the form $\rho(a,t)=A(a)T(t)$ and have shown that every age distribution approaches a persistent age distribution as $t\to\infty$. We will show that in the harvested case there may also be age distributions $\rho(a,t)$ which are identically zero for large t.

In the harvest case, where $\nu(\eta)\not\equiv 0$, the presence of the term in (25) which depends only on a suggests that the analogue of a persistent age distribution is a solution of the form

$$\rho(a,t)=A(a)T(t)-g(a), \tag{27}$$

where we may assume $\int_0^\infty A(a)\,da=1$ with no loss of generality. If we substitute the form (27) into (3), we obtain

$$A'(a)T(t)+A(a)T'(t)+\mu(a)A(a)T(t)=g'(a)+\mu(a)g(a)-\nu(a),$$

where primes denote differentiation with respect to either a or t. Since the right-hand side is independent of t,

$$\frac{d}{dt}[A'(a)T(t)+A(a)T'(t)+\mu(a)A(a)T(t)]$$
$$=A(a)T''(t)+\{A'(a)+\mu(a)A(a)\}T'(t)=0.$$

Treating this as a first order differential equation for $T'(t)$, we find

$$T'(t)=T'(0)\exp\left(-\left\{\frac{A'(a)}{A(a)}+\mu(a)\right\}t\right),$$

But since $T'(t)$ is independent of a, $A'(a)/A(a)+\mu(a)$ must be a constant $-k$. If $k\neq 0$, this implies $T'(t)=T'(0)e^{kt}$ or $T(t)=T(0)e^{kt}$ as well as $A(a)=A(0)\exp(-ka-\int_0^a\mu(\alpha)\,d\alpha)$. It follows that $A'(a)T(t)+A(a)T'(t)+\mu(a)A(a)T(t)=0$, and (27) becomes $g'(a)+\mu(a)g(a)-\nu(a)=0$, which yields

$$g(a)=\pi_0(a)\left[g(0)+\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right],$$

using (25). Combining the solutions for $A(a)$, $T(t)$ and $g(a)$, we obtain from (27)

$$\rho(a,t)=\pi_0(a)\left[ce^{k(t-a)}-g(0)-\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right].$$

Substitution of this into $\rho(0,t)=\int_0^\infty \beta(a)\rho(a,t)\,da$ gives, for $0\le t<\infty$

$$\begin{aligned} ce^{kt}-g(0)&=\int_0^\infty \beta(a)\pi_0(a)\left[ce^{k(t-a)}-g(0)-\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right]da\\ &=ce^{kt}\int_0^\infty \beta(a)\pi_0(a)e^{-ka}\,da-g(0)\int_0^\infty \beta(a)\pi_0(a)\,da\\ &\quad-\int_0^\infty \beta(a)\pi_0(a)\left[\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right]da\\ &=ce^{kt}\int_0^\infty \beta(a)\pi_0(a)e^{-ka}\,da-g(0)\int_0^\infty \beta(a)\pi_0(a)\,da-h_1(\infty), \end{aligned}$$

using $h_1(\infty)=\lim_{t\to\infty}h_1(t)=\int_0^\infty \beta(a)\pi_0(a)[\int_0^a(\nu(\eta)/\pi_0(\eta))\,d\eta]\,da$. From this relation, we have

$$\int_0^\infty \beta(a)\pi_0(a)e^{-ka}\,da=1, \tag{28}$$

$$g(0)\left[1-\int_0^\infty \beta(a)\pi_0(a)\,da\right]=h_1(\infty)\ge 0. \tag{29}$$

If $k>0$, (28) implies $\int_0^\infty \beta(a)\pi_0(a)\,da>1$, and then (29) implies $g(0)<0$. Similarly, if $k<0$, then $g(0)>0$.

The case $k=0$ must be treated slightly differently. In this case, $T'(t)=T'(0)$ and $T(t)=T'(0)t$, while $A(a)=A(0)\pi_0(a)$. Then

$$A'(a)T(t)+A(a)T'(t)+\mu(a)A(a)T(t)=A(0)T'(0)\pi_0(a),$$

and we have

$$g'(a)+\mu(a)g(a)-\nu(a)=A(0)T'(0)\pi_0(a).$$

This leads to

$$g(a)=\pi_0(a)\left[g(0)+\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta+aA(0)T'(0)\right].$$

Now (27) gives

$$\rho(a,t)=\pi_0(a)\left[c(t-a)-g(0)-\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right]$$

and substitution into $\rho(0,t)=\int_0^\infty \beta(a)\rho(a,t)\,da$ gives

$$\begin{aligned} ct-g(0)&=\int_0^\infty \beta(a)\pi_0(a)\left[c(t-a)-g(0)-\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}d\eta\right]da\\ &=ct\int_0^\infty \beta(a)\pi_0(a)\,da-c\int_0^\infty a\beta(a)\pi_0(a)\,da\\ &\quad-g(0)\int_0^\infty \beta(a)\pi_0(a)\,da-h_1(\infty). \end{aligned}$$

This implies

$$\int_0^\infty \beta(a)\pi_0(a)\,da=1, \tag{30}$$

$$g(0)=c\int_0^\infty a\beta(a)\pi_0(a)\,da+g(0)\int_0^\infty \beta(a)\pi_0(a)\,da+h_1(\infty). \tag{31}$$

Substitution of (30) into (31) gives $c\int_0^\infty a\beta(a)\pi_0(a)\,da+h_1(\infty)=0$. As $c\geq 0$, this is possible only if $c=0$, corresponding to an equilibrium age distribution, and $h_1(\infty)=0$, corresponding to zero harvest. Thus if the birth and death moduli are functions of age only and if there is harvesting, the only persistent age distributions correspond to populations which grow exponentially or which tend to zero exponentially in time.

To examine the asymptotic behaviour as $t\to\infty$ of an age distribution $\rho(a,t)$, we study the asymptotic behaviour of $B(t)$ as a solution of the first equation in (24) and then use (26). In order to do this, we write

$$B(t)=B_1(t)-B_2(t),$$

where $B_1(t)\geq 0$, $B_2(t)\geq 0$ and

$$B_1(t)=b(t)+\int_0^t \beta(a)\pi_0(a)B_1(t-a)\,da, \tag{32}$$

$$B_2(t)=h_1(t)+\int_0^t \beta(a)\pi_0(a)B_2(t-a)\,da. \tag{33}$$

Now (32) and (33) are linear renewal equations with $b(t)\geq 0$, $h_1(t)\geq 0$. We assume that the kernel $\beta(a)\pi_0(a)\geq 0$ satisfies

$$\int_0^\infty \beta(a)\pi_0(a)\,da<\infty, \qquad \int_0^\infty a\beta(a)\pi_0(a)\,da<\infty \tag{34}$$

and that there is harvesting, $\lim_{t\to\infty}h_1(t)>0$. To analyze (32), we assume that $\int_0^\infty b(a)\,da<\infty$ if $\int_0^\infty \beta(a)\pi_0(a)\,da\leq 1$ and that b is bounded on $0\leq a<\infty$ if $\int_0^\infty \beta(a)\pi_0(a)\,da>1$; this hypothesis is biologically plausible since b would normally have compact support. Then it is known (Feller [7]) that if $\int_0^\infty \beta(a)\pi_0(a)\,da<1$, then $\int_0^\infty B_1(t)\,dt<\infty$, and thence that $\lim_{t\to\infty}B_1(t)=0$, while if $\int_0^\infty \beta(a)\pi_0(a)\,da\geq 1$ and $p\geq 0$ is defined by

$$\int_0^\infty \beta(a)\pi_0(a)e^{-pa}\,da=1, \tag{35}$$

then $B_1(t)\sim c_1e^{pt}$ as $t\to\infty$.

To analyze (33), we differentiate to obtain

$$B_2'(t)=h_1'(t)+\int_0^t \beta(a)\pi_0(a)B_2'(t-a)\,da, \tag{36}$$

where

$$h_1'(t)=\int_0^\infty \frac{\nu(\eta)}{\pi_0(\eta)}\beta(\eta+t)\pi_0(\eta+t)\,d\eta\geq 0$$

and

$$\int_0^\infty h_1'(t)\,dt=h_1(\infty)=\int_0^\infty \beta(a)\pi_0(a)\left[\int_0^a \frac{\nu(\eta)}{\pi_0(\eta)}\,d\eta\,da\right].$$

We apply the results of [7] to the differentiated equation (36) to conclude that if $\int_0^\infty(a)\pi_0(a)\,da<1$ then $0<\int_0^\infty B_2'(t)\,dt<\infty$, whence $\lim_{t\to\infty}B_2(t)>0$, if $\int_0^\infty\beta(a)\pi_0(a)\,da=1$, then $\lim_{t\to\infty}B_2'(t)<\infty$, whence $B_2(t)\sim c_2 t$ as $t\to\infty$, while if $\int_0^\infty\beta(a)\pi_0(a)\,da>1$, then $B_2(t)\sim c_2 e^{pt}$ as $t\to\infty$, where p is again defined by (35). Combining the results for (32) and (33), we see that if $\int_0^\infty\beta(a)\pi_0(a)\,da\le 1$, then $B_1(t)-B_2(t)$ is negative for all sufficiently large t, which means that $B(t)$ reaches zero in finite time. On the other hand, if $\int_0^\infty\beta(a)\pi_0(a)\,da>1$, then $B(t)\sim(c_1-c_2)e^{pt}$ as $t\to\infty$. If $c_1>c_2$, then (26) shows that $\rho(a,t)$ approaches a persistent age distribution. If $c_1<c_2$, which more detailed study of [7] shows is equivalent to

$$\int_0^\infty e^{-pa}\{b(a)-h_1(a)\}\,da<0,$$

then $B(t)$ reaches zero in finite time just as in the case $\int_0^\infty\beta(a)\pi_0(a)\,da\le 1$ and $\rho(a,t)=0$ if $(t-a)$ is sufficiently large.

We may summarize as follows:

THEOREM 2. *Suppose the birth and death moduli are functions of age only, that* (34) *is satisfied, that* $\int_0^\infty b(a)\,da<\infty$ *if* $\int_0^\infty\beta(a)\pi_0(a)\,da\le 1$ *and that b is bounded on* $0\le a<\infty$ *if* $\int_0^\infty\beta(a)\pi_0(a)\,da>1$, *and that there is harvesting. Then every age distribution either vanishes identically for* $(t-a)$ *sufficiently large or approaches a persistent age distribution as* $t\to\infty$.

Theorem 2 says that the harvested situation differs from the unharvested situation in two respects. The population may die out in finite time, which cannot happen without harvesting, and there cannot be an equilibrium age distribution, which can happen without harvesting. In the harvested use the population either dies out in finite time or grows exponentially.

6. The situation in which the birth modulus is a function of age only, $\beta=\beta(a)$, with $\int_0^\infty\beta(a)\,da<\infty$ and the death modulus is a function of population size only, $\mu=\mu(P)$, has been examined by Gurtin and MacCamy [12]. In studying this situation, it is convenient to make a change of variables in integrals such as $\int_0^a\mu(P(t-a+\alpha))\,d\alpha$, to write them as $\int_{t-a}^t\mu(P(u))\,du$. With this change, the system (8), (14) takes the form

$$B(t)=b(t)+\int_0^t\beta(a)\exp\left(-\int_{t-a}^t\mu(P(u))\,du\right)B(t-a)\,da-h_1(t), \tag{37}$$

$$P'(t)=B(t)-P(t)\mu(P(t))-H, \tag{38}$$

where

$$H=\int_0^\infty\nu(a)\,da,$$

$$b(t)=\int_t^\infty\beta(a)\phi(a-t)\exp\left(-\int_0^t\mu(P(u))\,du\right)da,$$

$$h_1(t)=\int_0^\infty\nu(\eta)\left[\int_\eta^{n+t}\beta(a)\exp\left(-\int_{t-a+\eta}^t\mu(P(u))\,du\right)da\right]d\eta.$$

If we make the change of variables

$$B^*(t)=B(t)\exp\left(\int_0^t\mu(P(u))\,du\right),\qquad P^*(t)=P(t)\exp\left(\int_0^t\mu(P(u))\,du\right),$$

(37), (38) become

$$B^*(t)=b^*(t)+\int_0^t \beta(a)B^*(t-a)\,da-h_1^*(t), \tag{39}$$

$$P^{*\prime}(t)=B^*(t)-H+\exp\left(\int_0^t \mu(P(u))\,du\right), \tag{40}$$

with

$$\begin{aligned} b^*(t)&=\int_t^\infty \beta(a)\phi(a-t)\,da,\\ h_1^*(t)&=\int_0^\infty \nu(\eta)\left[\int_\eta^{n+t}\beta(a)\exp\left(\int_0^{t-a+\eta}\nu(P(u))\,du\right)da\right]d\eta. \end{aligned} \tag{41}$$

Much as in the preceeding section, we write

$$B^*(t)=B_1^*(t)-B_2^*(t)$$

with $B_1^*(t)\geq 0$, $B_2^*(t)\geq 0$ and

$$\begin{aligned} B_1^*(t)&=b^*(t)+\int_0^t \beta(a)B_1^*(t-a)\,da,\\ B_2^*(t)&=h_1^*(t)+\int_0^t \beta(a)B_2^*(t-a)\,da. \end{aligned}$$

We then conclude that if $\int_0^\infty \beta(a)\,da<1$, $\lim_{t\to\infty}B_1^*(t)=0$, if $\int_0^\infty \beta(a)\,da=1$, $\lim_{t\to\infty}B_1^*(t)<\infty$ and if $\int_0^\infty \beta(a)\,da>1$ and $p>0$ is defined by $\int_0^\infty \beta(a)e^{-pa}\,da=1$ then $B_1^*(t)\sim c_1e^{pt}$ as $t\to\infty$.

In order to obtain analogous results for $B_2^*(t)$, we must estimate $h_1^*(t)$. Differentiation of (41) under the integral sign gives

$$\begin{aligned} h_1^{*\prime}(t)&=\int_0^\infty \nu(\eta)\beta(\eta+t)\,d\eta\\ &\quad+\int_0^\infty \nu(\eta)\left[\int_\eta^{n+t}\beta(a)\exp\left(\int_0^{t-a+\eta}\mu(P(u))\,du\right)\mu(P(t-a+\eta))\,da\right]d\eta\\ &\geq\int_0^\infty \nu(\eta)\beta(n+t)\,dn\geq 0. \end{aligned}$$

Thus $h_1^*(t)$ is an increasing function. We now use the comparison principle of Nohel [20], which states that $B_2^*(t)$ is not less than the solution of the integral equation.

$$z(t)=\int_0^\infty \nu(\eta)\beta(\eta+t)\,d\eta+\int_0^t \beta(a)z(t-a)\,da,$$

where

$$\begin{aligned} \int_0^\infty\left[\int_0^\infty \nu(\eta)\beta(\eta+t)\,d\eta\right]dt&=\int_0^\infty\left[\int_0^\infty \beta(\eta+t)\,dt\right]\nu(\eta)\,d\eta\\ &=\int_0^\infty\left[\int_\eta^\infty \beta(a)\,da\right]\nu(\eta)\,d\eta\\ &\leq\int_0^\infty \nu(\eta)\,d\eta\int_0^\infty \beta(a)\,da<\infty. \end{aligned}$$

It follows, again, much as in the preceding section, that if $\int_0^\infty \beta(a)\,da<1$, $B_2^*(t)$ is bounded away from zero as $t\to\infty$, if $\int_0^\infty \beta(a)\,da=1$, $B_2^*(t)$ grows at least as rapidly as a constant multiple of t as $t\to\infty$ and if $\int_0^\infty \beta(a)\,da>1$, $B_2^*(t)\geq c_2 e^{pt}$ as $t\to\infty$ with $p>0$ defined by $\int_0^\infty \beta(a)e^{-pa}\,da=1$. Combining the estimates for $B_1^*(t)$ and $B_2^*(t)$ and recalling that

$$B(t)=B^*(t)\exp\left(-\int_0^t \mu(P(u))\,du\right)$$

$$=[B_1^*(t)-B_2^*(t)]\exp\left(-\int_0^t \mu(P(u))\,du\right),$$

we obtain the following analogue of Theorem 2.

THEOREM 3. *Suppose that the death modulus is a function of population size only, that the birth modulus is a function of age only, with* $\int_0^\infty \beta(a)\,da<\infty$, *and that there is harvesting. Then either the birth rate and population size are zero for all sufficiently large t or*

$$B^*(t)=B(t)\exp\left(\int_0^t \mu(P(u))\,du\right)\sim ce^{pt}$$

as $t\to\infty$, *with* $p>0$ *defined by* $\int_0^\infty \beta(a)e^{-pa}\,da=1$.

Theorem 3 provides information about the asymptotic behavior of $B(t)$ if $\int_0^\infty \beta(a)\,da\leq 1$ and if $\int_0^\infty \beta(a)\,da>1$, but the harvest rate is large enough that $B^*(t)$ becomes zero for all sufficiently large t. However, it leaves open the question of behavior of $B(t)$ and $P(t)$ if $\int_0^\infty \beta(a)\,da>1$ and $B^*(t)\sim ce^{pt}$ for some $c>0$, with $\int_0^\infty \beta(a)e^{-pa}\,da=1$. In order to treat this, we consider the three possibilities: (i) $\int_0^t \mu(P(u))\,du$ grows more slowly than pt as $t\to\infty$, (ii) $\int_0^t \mu(P(u))\,du$ grows more rapidly than pt as $t\to\infty$, (iii) $\lim_{t\to\infty}(1/t)\int_0^t \mu(P(u))\,du=P$. In the case (i), (43) implies that $P^{*\prime}(t)\sim ce^{pt}$ as $t\to\infty$, whence $P^*(t)\sim ce^{pt}$ and thus

$$P(t)\sim ce^{pt}\left(-\int_0^t \mu(P(u))\,du\right),$$

which is unbounded. Thus, by (38), $B(t)$ is also unbounded. In the case (ii), (40) implies that $P^{*\prime}(t)$ becomes ultimately negative and in fact that $P^*(t)$ must be zero for all sufficiently large t. (It is not difficult to show that this implies boundedness of $h_1^*(t)$.) Then $P(t)$ and $B(t)$ must be zero for all sufficiently large t. This, however, contradicts the hypothesis that $\int_0^t \mu(P(u))\,du$ grows more rapidly than pt unless $\mu(P)$ is a constant μ_0, in which case we must have $p<\mu_0$. In the case (iii), we conclude in a similar manner that $P(t)$ and $B(t)$ must grow more slowly than $e^{\varepsilon t}$ for any $\varepsilon>0$, but we cannot obtain more precise estimates. We can, however, conclude that $\rho(a,t)$ cannot tend to an equilibrium age distribution. If $\lim_{t\to\infty}B(t)=B_0$ and $\lim_{t\to\infty}P(t)=P_0$, then $\lim_{t\to\infty}(1/t)\int_0^t \mu(P(u))\,du=\mu(P_0)$ and thus $\mu(P_0)=p$. The equilibrium conditions (17), (18) imply $\int_0^\infty \beta(a)e^{-pa}\,da>1$ if there is a harvest, and this contradicts the definition of p. Our overall conclusion is less precise than may be hoped.

THEOREM 4. *Suppose that the death modulus is a function of population size only and that the birth modulus is a function of age only, with* $\int_0^\infty \beta(a)\,da<\infty$, *and that there is harvesting. Then either the birth rate and population size are zero for all sufficiently large t or they are unbounded or they do not tend to limits as* $t\to\infty$.

7. There are many directions in which improvements in our results would be desirable. For example, it would be of interest to determine whether in the case $\beta=\beta(a)$, $\mu=\mu(P)$ it is possible to have bounded solutions and to describe their asymptotic behavior if it is. We have not touched on the behavior of solutions if β and μ may depend on both age and population size, and the examples treated by Gurtin and MacCamy [12], [13] suggest that there are many different possibilities. We have also omitted any discussion of optimization of harvest and of nonconstant harvest rates. These topics will be the subject of further investigation.

REFERENCES

[1] F. BRAUER, *On a nonlinear equation for population growth problems*, this Journal, 6 (1975), pp. 312–317.

[2] _____, *Constant-rate harvesting of populations governed by Volterra integral equations*, J. Math. Anal. Appl., 56 (1976), pp. 18–27.

[3] _____, *Perturbations of the nonlinear renewal equation*, Adv. in Math., 22 (1976), pp. 32–51.

[4] K. L. COOKE AND J. A. YORKE, *Some equations modelling growth processes and gonorrhea epidemics*, Math Biosci., 16 (1973), pp. 75–101.

[5] J. M. CUSHING, *Lectures on Volterra integro-differential equations in population dynamics*, Proc. Centro Internazionale Mathematico Estivo Session on "Mathematics in Biology", Florence, Italy, 1979.

[6] _____, *Model stability and instability in age structured populations*, J. Theoret. Biol., 86 (1980), pp. 709–730.

[7] W. FELLER, *On the integral equations of renewal theory*, Ann. Math. Stat., 12 (1941), pp. 243–267.

[8] W. M. GETZ, *Optimal harvesting of structured populations*, Math. Biosci., 44 (1979), 269–291.

[9] _____, *The ultimate-sustainable-yield problem in nonlinear age-structured populations*, Math. Biosci., 48 (1980), pp. 279–292.

[10] D. H. GRIFFEL, *Age-dependent population growth*, J. Inst. Maths. Applics., 17 (1976), pp. 141–152.

[11] M. E. GURTIN AND R. C. MACCAMY, *Non-linear age-dependent population dynamics*, Arch. Rational Mech. Anal., 3 (1974), pp. 281–300.

[12] _____, *Population dynamics with age dependence*, Nonlinear Analysis & Mechanics: Herriat-Watt Symposium, Vol III, R. J. Knops, ed., Pitman, London, 1972, pp. 1–35.

[13] _____, *Some simple models for nonlinear age-dependent population dynamics*, Math. Biosci. 43 (1979), pp. 199–211.

[14] M. E. GURTIN AND L. F. MURPHY, *On optimal harvesting with an application to age-structured populations*, J. Math. Biol., 13 (1981), pp. 131–148.

[15] _____, *On the optimal harvesting of age-structured populations*, Differential Equations and Applications in Ecology, Epidemics, and Population Problems, S. N. Busenberg and K. L. Cooke, eds., Academic Press, New York, 1981, pp. 115–129.

[16] F. C. HOPPENSTEADT, *Mathematical Theories of Populations: Demographics, Genetics, and Epidemics.*, CBMS Regional Conference Series in Applied Mathematics 20, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[17] S. O. LONDEN, *On the solutions of a nonlinear Volterra equation*, J. Math. Anal. Appl., 39 (1972), pp. 564–573.

[18] _____, *On a non-linear Volterra integral equation*, J. Differential Equations, 14 (1973), pp. 106–120.

[19] A. G. MACKENDRICK, *Applications of mathematics to medical problems*, Proc. Edinburgh Math. Soc., 40 (1926), pp. 98–130.

[20] J. A. NOHEL, *Some problems in nonlinear Volterra integral equations*, Bull. Amer. Math. Soc., 68 (1962), pp. 323–329.

[21] C. RORRES, *Stability of an age specific population with density dependent fertility*, Theoret. Pop. Biol., 10 (1976), pp. 26–46.

[22] _____, *A nonlinear model of population growth in which fertility is dependent on birth rate*, SIAM J. Appl. Math.,37 (1979), pp. 423–432.

[23] _____, *Local stability of a population with density-dependent fertility*, Theoret. Pop. Biol., 16 (1979), pp. 283–300.

[24] C. RORRES AND W. FAIR, *Optimal age-specific harvesting policy for a continuous-time population model*, Modelling and Differential Equations in Biology, T. A. Burton, ed., Dekker, New York, 1980, pp. 239–254.

[25] D. A. SÁNCHEZ, *Linear age-dependent population growth with harvesting*, Bull. Math. Biol., 40 (1978), pp. 377–385.

[26] _____, *Linear age-dependent population growth with seasonal harvesting*, J. Math. Biol., 9 (1980), pp. 361–368.

[27] F. R. SHARPE AND A. J. LOTKA, *A problem in age distribution*, Philosophical Mag., 21 (1911), pp. 435–438.

[28] E. SINESTRARI, *Non-linear age-dependent population growth*, J. Math. Biol., 9 (1980), pp. 331–345.

[29] K. E. SWICK, *A model of single species population growth*, this Journal, 7 (1976), pp. 565–576.

[30] _____, *A nonlinear age-dependent model of a single species population dynamics*, SIAM J. Appl. Math., 32 (1977), pp. 484–498.

[31] _____, *Periodic solutions of a nonlinear age-dependent model of single species population dynamics*, this Journal, 11 (1980), pp. 901–910.

[32] _____, *A nonlinear model for human population dynamics*, SIAM J. Appl. Math., 40 (1981), pp. 266–278.

[33] H. VON FOERSTER, *Some remarks on changing populations*, The Kinetics of Cellular Proliferation, F. Stohlman, ed., Shalton, New York, 1959, pp. 382–407.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0010 $01.25/0

ON THE RATE OF CONVERGENCE OF STOPPED RANDOM WALKS*

JED CHAPIN† AND JOHN L. B. GAMLEN†

Abstract. Let $F_{x,t}(z)$ be the distribution function at time t of Brownian motion starting at $x \in [0, 1]$ with absorbing boundary points 0 and 1. For

$$t \in \left\{0, \frac{1}{2n^2}, \cdots, \frac{N}{2n^2}\right\} \quad \text{and} \quad x \in \left\{0, \frac{1}{n}, \cdots, \frac{k}{n}, \cdots, 1\right\},$$

let $F_{x,t}^{(n)}(z)$ be the distribution function at time t of the process obtained by stopping the standard random walk process at the boundaries 0 and 1. (Standard random walk is symmetric and has variance $1/n^2$ at $t = 1/2n^2$.) We prove that

$$\left|F_{x,t}^{(n)}(z) - F_{x,t}(z)\right| \leq \frac{3.6}{n\sqrt{t}} = 3.6\sqrt{2}\,\frac{1}{\sqrt{N}},$$

where $n \geq 10$, $t \geq \frac{1}{40}$ and N is even.

The Berry and Esseen estimate can be used to give the rate of convergence of distribution functions of unconfined random walks to the distribution functions of Brownian motion. Our work extends this result.

Introduction. The unstopped version of our result gives the well-known rate of convergence in the central limit problem for the usual random walks on $\mathbb{R}$. Our first step was to take this known result (which is a particular case of the Berry-Essen estimate [1, p. 206]) and formulate it in terms of stochastic processes, rather than individual distributions. We then proved the stopped version of this result.

Let $F_{x,t}(\cdot)$ be the distribution function at time t of Brownian motion starting at $x \in [0, 1]$, with absorbing endpoints 0, 1. Write

$$F_{x,t}^{(m)}(\cdot), \qquad t \in \left\{0, \frac{1}{2m^2}, \cdots, \frac{k}{2m^2}, \cdots\right\}, \quad x \in \left\{0, \frac{1}{m}, \frac{2}{m}, \cdots, 1\right\},$$

for the distribution function at time t of the process obtained by stopping "simple random walk" at the boundaries 0, 1. Here "simple random walk" means the random walk jumping at the above times left or right between points of the set $0, \frac{1}{m}, \frac{2}{m}, \cdots$. Later these processes will be more precisely described. We prove that; if m is even, x and z are even multiples of $\frac{1}{m}$ and t is an even multiple of $1/2m^2$; then

$$\left|F_{x,t}(z) - F_{x,t}^{(n)}(z)\right| \leq 3.6\,\frac{1}{m\sqrt{t}}.$$

One reason for tackling this problem is that in the Berry–Esseen result for the central limit theorem, random walk is the worst possible case, in the sense that the predicted rate of convergence holds exactly for random walk. One might reasonably hope that our problem gives the worst possible rate of convergence, among those problems which arise by stopping more general sums of independent identically distributed random variables. This amounts to the hope that the Berry–Esseen estimate extends to stopped sums of random variables.

It would be of some importance if our result could be extended to cover convergence of more general discrete processes to diffusions with nonconstant covariance velocity. Such problems arise in both genetics and in finite difference methods for

* Received by the editors October 13, 1981, and in revised form May 7, 1982.
† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

solving partial differential equations. For the Wright–Fisher discrete and continuous processes of genetics, S. Ethier and F. Norman [6] discuss rate of convergence of the expectations of C^4 functions. Applied to convergence of distribution functions, their theorem yields a slow rate of convergence. If our result is extended to their situation, it would be quite useful to geneticists, who use the Wright–Fisher diffusion to calculate approximate results for the Wright–Fisher discrete processes.

A greal deal has been written on convergence of finite-difference methods to solutions of differential equations, but not much of it is relevant for this particular problem, even though our problem is in disguise a finite-difference problem. However, the fact that convergence occurs in our problem was proved by P. Lax [5]. No rate of convergence could be extracted from the proof, unfortunately.

In §1 we transform our problem to a very similar one involving difference operators. The discrete and continuous distribution functions are then expressed as the solutions to boundary value problems, and then as discrete and continuous eigenfunction expansions.

In §2 the terms of the eigenfunction expansions are compared, and an estimate is obtained relating the continuous process to the modified discrete processes considered.

In §3 the main result is proved.

1. Eigenfunction expansion of discrete and continuous distribution functions. Consider $F_{x,t}(z)$, the distribution function at time t of Brownian motion starting at $x\in[0,1]$ and absorbed at the boundary points $0,1$. For z fixed but arbitrary in $[0,1)$ write $\phi(x,t)=F_{x,t}(z)-(1-x)$. From the backward equation for $F_{x,t}(z)$ we find that $\phi(x,t)$ is the unique solution to the boundary value problem

$$(*)\qquad \frac{\partial}{\partial t}\phi(x,t)=\frac{\partial^2\phi(x,t)}{\partial x^2},\qquad 0<x<1,\quad t>0,$$

$$\phi(0,t)=\phi(1,t)=0,$$

$$\phi(x,0)=\begin{cases} x & \text{if } x\le z,\\ x-1 & \text{otherwise.}\end{cases}$$

The solution may be obtained as the eigenfunction expansion:

$$(**)\qquad \phi(x,t)=\sum_{j=1}^{\infty}\hat{\phi}_j(t)Y_j(x),$$

where $Y_j(x)=\sqrt{2}\,\sin j\pi x$ and $\hat{\phi}_j(t)=\int_0^1\phi(x,t)Y_j(x)\,dx$. Integration by parts in $(*)$ yields that $\hat{\phi}_j(t)$ satisfies the initial value problem:

$$\lambda_j\hat{\phi}_j(t)=\frac{d}{dt}\hat{\phi}_j(t)\quad\text{with } \hat{\phi}_j(0)\text{ given,}$$

where the λ_j's are the eigenvalues associated with the Y_j's. The Y_j's are the normalized (in $L^2[0,1]$) eigenfunctions for the operator d^2/dx^2 on

$$\{f\in L^2[0,1]\,|\,f''\in L^2[0,1], f(0)=f(1)=0\}.$$

The series $(**)$ is well known to converge uniformly on $[0,1]$ for fixed $t>0$.

We will now write out some well-known computations (cf. [5, p. 65]) to aid us in their imitation in the discrete case.

To obtain the initial value problem, we integrate both sides of the equation (*) against Y_j, i.e.,

$$\frac{\partial}{\partial t}\int_0^1 \phi(x,t)Y_j(x)\,dx=\int_0^1 \frac{\partial^2\phi(x,t)}{\partial x^2}Y_j(x)\,dx.$$

Integrating by parts twice gives

$$\frac{\partial}{\partial t}\hat{\phi}_j(t)=\left[y_j(x)\frac{\partial\phi(x,t)}{\partial x}-\phi(x,t)\frac{\partial Y_j(x)}{\partial x}\right]_0^1+\int_0^1\phi(x,t)Y_j''(x)\,dx.$$

Observing that the boundary terms are zero and using $Y_j''(x)=\lambda_j Y_j(x)$, we get the stated initial value problem. The solution to the initial value problem is $\hat{\phi}_j(t)=\hat{\phi}_j(0)e^{\lambda_j t}$. To find $\hat{\phi}_j(0)$, we calculate

$$\begin{aligned}\int_0^1\phi(x,0)Y_j(x)\,dx&=\frac{1}{\lambda_j}\left\{\int_0^z\phi(x,0)Y_j''(x)\,dx+\int_z^1\phi(x,0)Y_j''(x)\,dx\right\}\\&=\frac{1}{\lambda_j}\left\{\left[\phi(x,0)Y_j'(x)-Y_j\frac{\partial\phi(x,0)}{\partial x}\right]_0^z\right.\\&\qquad+\left[\phi(x,0)Y_j'(x)-Y_j\frac{\partial\phi(x,0)}{\partial x}\right]_z^1\\&\qquad\left.+\int_0^z\frac{\partial^2\phi(x,0)}{\partial x^2}Y_j(x)\,dx+\int_z^1\frac{\partial^2\phi(x,0)}{\partial x^2}Y_j(x)\,dx\right\}.\end{aligned}$$

Thus $\hat{\phi}_j(0)=Y_j'(z)/\lambda_j=-\cos j\pi z/j\pi$.

We now imitate the above constructions in our discrete situation. Recall from the introduction the distribution functions:

$$F_{x,t}^{(m)}(z),\qquad z,x\in\left\{0,\frac{1}{m},\cdots,\frac{k}{m},\cdots,1\right\},\quad t\in\left\{0,\frac{1}{2m^2},\frac{2}{2m^2},\cdots\right\}.$$

These are the distribution functions associated with the transition functions:

$$P_t^{(m)}(x,y)=\begin{cases}1 & \text{if } x=y \quad (x=0 \text{ or } x=1),\\ \frac{1}{2} & \text{if } |y-x|=\frac{1}{m},\quad x\neq 0,\quad x\neq 1,\\ 0 & \text{otherwise,}\quad \text{for } t=\frac{1}{2m^2}.\end{cases}$$

By modifying these processes we contrive to deal with second order difference operators as generators. Precisely, define transition functions.

$$Q_t^{(n)}(x,y)=\begin{cases}1 & \text{if } x=y \quad (x=0 \text{ or } x=1),\\ \frac{1}{2} & \text{if } x=y,\quad x\neq 0,\quad x\neq 1,\\ \frac{1}{4} & \text{if } |y-x|=\frac{1}{n},\quad x\neq 0,\quad x\neq 1,\\ 0 & \text{otherwise,}\quad \text{for } t=\frac{1}{4n^2}.\end{cases}$$

Its distribution functions $G_{x,t}^{(n)}(\cdot)$ satisfy:

$$G_{x,t}^{(n)}(z)=F_{x,t}^{(2n)}(z),$$

where

$$x,z\in\left\{0,\frac{1}{n},\frac{2}{n},\cdots,1\right\},\qquad t\in\left\{0,\frac{1}{4n^2},\frac{2}{4n^2},\cdots\right\}.$$

Define the (standard) difference operators Δ_x, Δ_t and δ_x^2 as follows:

$$\Delta_x g(x,t)=g\left(x+\frac{1}{n}\right)-g(x)\quad\text{(similarly for }\Delta_t),$$

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

Write $\phi^{(n)}(x,t)=G_{x,t}^{(n)}(z)-(1-x)$. Then $\phi^{(n)}$ is the unique solution to the discrete B.V.P.

$$\Delta_t\phi^{(n)}(x,t)\div\left(\frac{1}{4}n^2\right)=\delta^2\phi^{(n)}(x,t)\div\left(\frac{1}{n}^2\right),\qquad x\in\left\{\frac{1}{n},\frac{2}{n},\cdots,\frac{n-1}{n}\right\},\quad t\in\left\{0,\frac{1}{4n^2},\frac{2}{4n^2},\cdots\right\},$$

$$\phi^{(n)}(0,t)=\phi^{(n)}(1,t)=0,$$

$$\phi^{(n)}(x,0)=\begin{cases}x & \text{if } x\le z,\\ x-1 & \text{if } x>z.\end{cases}$$

Let $Y_1^{(n)}\cdots Y_{n-1}^{(n)}$ be the normalized eigenfunctions of δ^2, and let $\lambda_1^{(n)}\cdots\lambda_{n-1}^{(n)}$ be the associated eigenvalues. We claim that the solution to the above B.V.P. is:

$$\phi^{(n)}(x,t)=\frac{1}{n}\sum_{j=1}^{n-1}\hat{\phi}_j^{(n)}(t)Y_j^{(n)}(x).$$

This may be verified directly, as in the continuous case. However, it may be derived using the discrete analogue of the eigenfunction transform technique, using summation by parts twice.

$$\sum_{k=1}^{n-1}p_k\delta^2 q_k=p_n\Delta q_{n-1}-p_1\Delta q_0-q_n\Delta p_{n-1}+q_1\Delta p_0+\sum_{k=1}^{n-1}q_k\delta^2 p_k.$$

Next we find the $Y_j^{(n)}$'s explicitly; then we find $\hat{\phi}_j(t)$, completing the explicit solution for $\phi^{(n)}(x,t)$.

$Y_j^{(n)}(\cdot)$ and $\{\lambda_j^{(n)}\}_{j=1}^{n-1}$ must satisfy

$$\frac{\Delta_x^2 Y_j^{(n)}(x_{k-1})}{1/n^2}=\lambda_j^{(n)}Y_j^{(n)}(x_k),\qquad Y_j^{(n)}(0)=Y_j^{(n)}(1)=0$$

and

$$\sum_{k=1}^{n-1}Y_j^{(n)^2}(x_k)\frac{1}{n}=1.$$

Recall the identities

$$\sin(a+b)-\sin(a-b)=2\sin b\cos a,$$
$$\cos(a+b)-\cos(a-b)=2\sin b\sin a.$$

These yield

$$\Delta\sin ux_k=2\sin\frac{u}{2n}\cos u\left(x_k+\frac{1}{2n}\right),$$
$$\Delta\cos ux_k=-2\sin\frac{u}{2n}\sin u\left(x_k+\frac{1}{2n}\right).$$

Iteration gives

$$\Delta^2\sin ux_{k-1}=-4\sin^2\frac{u}{2n}\sin ux_k,$$
$$\Delta^2\cos ux_{k-1}=-4\sin^2\frac{u}{2n}\cos ux_k.$$

Thus if we let $Y_j^{(n)}(x)=\sin j\pi x$ it would satisfy almost all of the required conditions. In that case

$$\lambda_j^{(n)}=-4n^2\sin^2\frac{j\pi}{2n}=-2n^2\left(1-\cos\frac{j\pi}{n}\right).$$

To normalize, consider

$$\sum_{k=1}^{n-1}\sin^2(j\pi x_k)\frac{1}{n}=\sum_{k=1}^{n-1}\frac{1}{2}(1-\cos 2j\pi x_k)\frac{1}{n}$$
$$=\frac{1}{2}-\frac{1}{2n}-\sum_{k=1}^{n-1}\cos(2j\pi x_k)\frac{1}{n}.$$

Then

$$\sum_{k=1}^{n-1}\cos(2j\pi x_k)\frac{1}{n}=\frac{1}{\lambda_{2j}^{(n)}}\sum_{k=1}^{n-1}\frac{\Delta^2\cos(2j\pi x_{k-1})}{(1/n)^2}\frac{1}{n}$$
$$=\frac{1}{\lambda_{2j}^{(n)}}\left[\frac{\Delta\cos(2j\pi x_k)}{1/n}\right]_{k=0}^{k=n-1}$$
$$=\frac{1}{\lambda_{2j}^{(n)}}\left(\frac{2\Delta\cos(0)}{1/n}\right)=\frac{1}{\lambda_{2j}^{(n)}}2n\left(\cos\frac{2j\pi}{n}-1\right).$$

Recall $\lambda_{2j}^{(n)}=-2n^2(1-\cos 2j\pi/n)$, so $\sum_{k=1}^{n-1}\cos(2j\pi x_k)/n=1/n$. Now we have

$$\sum_{k=1}^{n-1}\sin^2(j\pi x_k)\frac{1}{n}=\frac{1}{2}-\frac{1}{2n}+\frac{1}{n}=\frac{1+1/n}{2}.$$

Now we can write

$$Y_j^{(n)}(x)=\sqrt{\frac{2}{1+1/n}}\sin j\pi x.$$

It is easily verified that the solution to the discrete initial value problem (****) is

$$\hat{\phi}_j^{(n)}(t)=\hat{\phi}_j^{(n)}(0)\left(1+\lambda_j^{(n)}\cdot\frac{1}{4n^2}\right)^{4n^2t}$$

To find $\hat{\phi}_j^{(n)}(0)=\sum_{k=1}^{n-1}\phi^{(n)}(x_k,0)Y_j^{(n)}(x_k)\cdot 1/n$, use summation by parts, obtaining

$$\hat{\phi}_j^{(n)}(0)=\frac{(1+1/n)\Delta Y_j^{(n)}(z)}{\lambda_j^{(n)}\cdot 1/n}.$$

This complicated way of displaying the right-hand side will be useful when we compare with $\hat{\phi}_j(0)$ in the next section.

2. Comparison of eigenfunction expansions of discrete and continuous distribution functions. Now we are ready to compare

$$\phi(x,t)=\sum_{j=1}^{\infty}\hat{\phi}_j(0)e^{\lambda_j t}\sqrt{2}\sin j\pi x$$

with

$$\phi^{(n)}(x,t)=\sum_{j=1}^{n-1}\hat{\phi}_j(0)\left(1+\frac{\lambda_j^{(n)}}{4n^2}\right)^{4n^2t}\sqrt{\frac{2}{1+1/n}}\sin j\pi x.$$

In order to make the comparison we change the form of the expressions. Observe $\lambda_j t=-(j\pi)^2 t=-(j\pi/2n)^2 4n^2 t$. Also,

$$1+\frac{\lambda_j^{(n)}}{4n^2}=1-4n^2\left(\sin^2\frac{j\pi}{2n}\right)\frac{1}{4n^2}=\cos^2\frac{j\pi}{2n}.$$

Thus

$$\phi(x,t)=\sum_{j=1}^{\infty}\hat{\phi}_j(0)\left[e^{-(j\pi/2n)^2}\right]^{4n^2t}\sqrt{2}\sin j\pi x,$$

$$\phi^{(n)}(x,t)=\sum_{j=1}^{n-1}\hat{\phi}_j^{(n)}(0)\left[\cos^2\frac{j\pi}{2n}\right]^{4n^2t}\sqrt{\frac{2}{1+1/n}}\sin j\pi x.$$

This last expression is related to others in the literature (cf. [3, p. 353]). To clarify the overall scheme for estimating $|\phi^{(n)}(x,t)-\phi(x,t)|$, we denote the various quantities as shown

$$a_j=\hat{\phi}_j(0),\quad b_j^{(n)}=\left[e^{-(j\pi/2n)^2}\right]^{4n^2t},\quad c_j=\sqrt{2}\sin j\pi x,$$

$$a_j^{(n)}=\hat{\phi}_j^{(n)}(0),\quad b_j^{(n)}=\left[\cos^2\frac{j\pi}{2n}\right]^{4n^2t},\quad c_j^{(n)}=\sqrt{\frac{2}{1+1/n}}\sin j\pi x.$$

We will deal with the initial parts and the tails of the sums separately, writing

$$(*)\qquad \left|\sum_{j=1}^{n-1}a_j^{(n)}b_j^{(n)}c_j^{(n)}-\sum_{j=1}^{\infty}a_jb_jc_j\right|$$

$$\le\left|\sum_{1\le j\le n/2}a_j^{(n)}b_j^{(n)}c_j^{(n)}-a_jb_jc_j\right|$$

$$+\left|\sum_{n/2\le j\le n-1}a_j^{(n)}b_j^{(n)}c_j^{(n)}\right|+\left|\sum_{n/2<j<\infty}a_jb_jc_j\right|.$$

Let $\|a\| = \max_j |a_j|$, and let $\|c\|$, $\|a^{(n)}\|$, $\|a^{(n)} - a\|$ and $\|c^{(n)} - c\|$ be defined in a corresponding manner. Next we observe

$$(**) \quad \left| \sum_{1 \le j \le n/2} a_j^{(n)} b_j^{(n)} c_j^{(n)} - a_j b_j c_j \right|$$

$$\le \|a^{(n)} - a\| \sum_{1 \le j \le n/2} |b_j| \|c\|$$

$$+ \|a\| \sum_{1 \le j \le n/2} |b_j^{(n)} - b_j| \|c\| + \|a\| \sum_{1 \le j \le n/2} |b_j| \|c^{(n)} - c\|$$

$$+ \|a\| \sum_{1 \le j \le n/2} |b_j^{(n)} - b_j| \|c^n - c\| + \|a^{(n)} - a\| \sum_{1 \le j \le n/2} |b_j| \|c^{(n)} - c\|$$

$$+ \|a^{(n)} - a\| \sum_{1 \le j \le n/2} |b_j^{(n)} - b_j| \|c\| + \|a^{(n)} - a\| \sum_{1 \le j \le n/2} |b_j^{(n)} - b_j| \|c^{(n)} - c\|.$$

This deals with the first term of (*).

For the second term of (*), write:

$$(***) \quad \left| \sum_{n/2 < j \le n-1} a_j^{(n)} b_j^{(n)} c_j^{(n)} \right| \le (\|a\| + \|a^{(n)} - a\|) \cdot (\|c\| + \|c^{(n)} - c\|) \sum_{n/2 < j \le n-1} |b_j^{(n)}|.$$

For the third term of (*), write

$$(****) \quad \left| \sum_{n/2 < j < \infty} a_j b_j c_j \right| \le \|a\| \|c\| \sum_{n/2 < j < \infty} |b_j|.$$

We will prove that for $n \ge 10$, $t \ge 1/4n$, we have

$$\left| \sum_{j=1}^{n-1} a_j b_j c_j - \sum_{j=1}^{\infty} a_j b_j c_j \right| \le \frac{1.8}{n\sqrt{t}}.$$

To do this we need only (**), (***), (****) together with the following estimates, whose proofs occupy the rest of the paper.

Estimates to be proved:

1. $\displaystyle\sum_{n/2 < j < n-1} |b_j^{(n)}| \le .0078 \, \frac{1}{n\sqrt{t}}$;

2. $\displaystyle\sum_{n/2 < j < \infty} |b_j| \le .0157 \, \frac{1}{n\sqrt{t}}$;

3. $\|a\| \le .4502$;

4. $\displaystyle\sum_{1 \le j \le n/2} |b_j| \le .2821 \, \frac{1}{\sqrt{t}}$;

5. $\|c\| \le \sqrt{2}$;

6. $\|a^{(n)} - a\| \le 2.037 \, \frac{1}{n}$;

7. $\displaystyle\sum_{1 \le j \le n/2} |b_j^{(n)} - b_j| \le .8031 \, \frac{1}{n\sqrt{t}}$;

8. $\|c^{(n)} - c\| \le \frac{1}{\sqrt{2}} \, \frac{1}{n}$.

For the first estimate note that $b_j^{(n)}=[\cos^2(j\pi/2n)]^{4n^2t}$ is decreasing in j, and that $|b_j^{(n)}|\le\cos^{8n^2t}(\pi/4)$, for $-n/2<j<n-1$. Thus

$$\sum_{n/2\le j\le n-1}|b_j^{(n)}|\le\frac{n}{2}\left(\frac{1}{\sqrt{2}}\right)^{8n^2t}\le .0078\,\frac{1}{n\sqrt{t}}.$$

For the second estimate,

$$\sum_{n/2<j<\infty}|b_j|\le\sum_{n/2<j<\infty}e^{-(j\pi)^2t}\le\int_{n/2}^{\infty}e^{-([y-1]\pi)^2t}\,dy$$

$$\le\frac{1}{\pi\sqrt{2t}}\int_{2\pi n\sqrt{2t}/5}^{\infty}e^{-u^2/2}\,du\le\frac{1}{\pi\sqrt{2t}}\,\frac{e^{-4\pi^2n^2t/25}}{2\pi n\sqrt{2t}/5},$$

$u=(y-1)\pi\sqrt{2t}$ (the last inequality follows from [6, p. 4])

$$\le\frac{5\sqrt{10}}{2\pi^2}e^{-2\pi^2/5}\,\frac{1}{n\sqrt{t}}\le(.0157)\,\frac{1}{n\sqrt{t}}.$$

For the third estimate,

$$|a_j|=|\hat\phi_j(0)|=\left|\frac{-\sqrt{2}\cos z}{j\pi}\right|\le\frac{\sqrt{2}}{\pi}\le .4502,$$

hence $\|a\|\le .4502$.

For the fourth estimate note that

$$\sum_{1\le j\le n/2}|b_j|\le\sum_{j=1}^{\infty}e^{-(j\pi)^2t}\le\int_0^{\infty}e^{-(y\pi)^2t}\,dy$$

$$=\frac{1}{\pi\sqrt{2t}}\int_0^{\infty}e^{u^2/2}\,du=\frac{1}{2\sqrt{\pi t}}\le .2821\,\frac{1}{\sqrt{t}}$$

$(u=\pi\sqrt{2t}\,y)$.

For the fifth estimate

$$|c_j|=\left|\sqrt{2}\sin j\pi x\right|\le\sqrt{2},$$

so $\|c\|\le\sqrt{2}$.

The sixth estimate is more technical. To find $\|a^{(n)}-a\|$ we need to compare

$$\hat\phi_j(0)=\frac{-\sqrt{2}}{j\pi}\cos j\pi z\quad\text{with}\quad\hat\phi_j^{(n)}(0)=\frac{1+(1/n)\Delta Y_j^{(n)}(z)}{\lambda_j^{(n)}(1/n)}.$$

$$\frac{\Delta Y_j^{(n)}(z)}{1/n}=\sqrt{\frac{2}{1+1/n}}\,n\left(\sin\left(j\pi z+\frac{j\pi}{n}\right)-\sin j\pi z\right)$$

$$=\sqrt{\frac{2}{1+1/n}}\,n\left(\sin\frac{j\pi}{n}\cos j\pi z+\sin j\pi z\left(\cos\frac{j\pi}{n}-1\right)\right).$$

We have

$$\lambda_j^{(n)}=2n^2\left(\cos\frac{j\pi}{n}-1\right)=-4n^2\sin^2\left(\frac{j\pi}{2n}\right).$$

So

$$\frac{1}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{1/n}=\sqrt{\frac{2}{1+1/n}}\left(\frac{n\sin(j\pi/n)}{-4n^2\sin^2(j\pi/2n)}\cos j\pi z+\frac{1}{2n}\sin j\pi z\right).$$

At this point we need a lemma.

LEMMA.

$$\left|\frac{1}{j\pi}-\frac{n\sin(j\pi/n)}{4n^2\sin^2(j\pi/2n)}\right|\leq\frac{1}{\pi}\frac{1}{n}\quad \textit{for } 1\leq j\leq n.$$

Proof. We have

$$\begin{aligned}\left|\frac{1}{j\pi}-\frac{n\sin(j\pi/n)}{4n^2\sin^2(j\pi/2n)}\right|&=\left|\frac{1}{j\pi}-\frac{\cos(j\pi/2n)}{2n\sin(j\pi/2n)}\right|\\&=\left|\frac{1}{j\pi}-\frac{1}{j\pi}\frac{(j\pi/2n)\cos(j\pi/2n)}{\sin(j\pi/2n)}\right|\\&=\frac{1}{j\pi}\left|\frac{\sin(j\pi/2n)-(j\pi/2n)\cos(j\pi/2n)}{\sin(j\pi/2n)}\right|\\&=\frac{1}{n}\frac{1}{2}\left|\frac{\sin(j\pi/2n)-(j\pi/2n)\cos(j\pi/2n)}{(j\pi/2n)\sin(j\pi/2n)}\right|.\end{aligned}$$

Now use the result that $0\leq(\sin x-x\cos x)/x\sin x\leq 2/\pi$ for $0<x\leq\pi/2$. So we get

$$\left|\frac{1}{j\pi}-\frac{n\sin(j\pi/n)}{4n^2\sin^2(j\pi/2n)}\right|\leq\frac{1}{n\pi}.\qquad\square$$

We now can proceed with our error estimate for $\|a^{(n)}-a\|$. Let

$$\varepsilon_1=\sqrt{\frac{2}{1+1/n}}-\sqrt{2},\qquad \varepsilon_2=\frac{1}{j\pi}-\frac{n\sin(j\pi/n)}{4n^2\sin^2(j\pi/2n)},\quad \varepsilon_3=\frac{1}{2n}\sin j\pi z.$$

We have

$$|\varepsilon_1|\leq\frac{1}{\sqrt{2}}\frac{1}{n},\quad |\varepsilon_2|\leq\frac{1}{\pi n},\quad |\varepsilon_3|\leq\frac{1}{2n}.$$

So

$$\begin{aligned}\frac{1}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{1/n}&=\left(\sqrt{2}+\varepsilon_1\right)\left[\left(-\frac{1}{j\pi}+\varepsilon_2\right)\cos j\pi z+\varepsilon_3\right]\\&=\frac{-\sqrt{2}\cos j\pi z}{j\pi}+\sqrt{2}\,\varepsilon_2\cos j\pi z+\sqrt{2}\,\varepsilon_3-\frac{\varepsilon_1}{j\pi}\cos j\pi z+\varepsilon_1\varepsilon_2\cos j\pi z+\varepsilon_1\varepsilon_3.\end{aligned}$$

Thus

$$\left|\hat{\phi}_j(0)-\frac{1}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{1/n}\right|\le\frac{\sqrt{2}}{\pi}\frac{1}{n}+\frac{1}{\sqrt{2}}\frac{1}{n}+\frac{1}{\sqrt{2}\,\pi n}+\frac{1}{\sqrt{2}\,\pi n^2}+\frac{1}{2\sqrt{2}\,n^2}$$

$$\le 1.441\,\frac{1}{n}.$$

Finally,

$$\left|\hat{\phi}_j(0)-\frac{1+1/n}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{1/n}\right|\le\left|\hat{\phi}_j(0)-\frac{1}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{Y_n}\right|+\frac{1}{n}\left|\hat{\phi}_j(0)\right|$$

$$+\frac{1}{n}\left|\hat{\phi}_j(0)-\frac{1}{\lambda_j^{(n)}}\frac{\Delta Y_j^{(n)}(z)}{Y_n}\right|$$

$$\le 2.037\,\frac{1}{n}.$$

This completes the proof of the sixth estimate.

For the seventh estimate we now find a bound for $\Sigma_{1\le j\le n/2}|b_j^{(n)}-b_j|$, which equals

$$\sum_{1\le j\le n/2}\left|\left[\cos^2\frac{j\pi}{2n}\right]^{4n^2t}-\left[e^{-(j\pi/2n)^2}\right]^{4n^2t}\right|.$$

The factorization $p^k-q^k=(p-q)\Sigma_{l=0}^{k-1}p^{k-1-l}q^l$ gives the formula

$$|p^k-q^k|\le|p-q|k\max\left\{|p|^{k-1},|q|^{k-1}\right\}.$$

Also,

$$\cos^2\theta=1-\theta^2+\frac{1}{3}\theta^4-\frac{2}{45}\theta^6\cos 2\xi\qquad(0<\xi<\theta)$$

and

$$e^{-\theta^2}=1-\theta^2+\frac{1}{2}\theta^4-\frac{1}{6}\theta^6e^{-\xi^2}\qquad(0<\xi<\theta).$$

So, for $0<j\pi/2n<\pi/4$,

$$\left|\cos^2\frac{j\pi}{2n}-e^{-(j\pi/2n)^2}\right|\le\frac{2}{7}\left(\frac{j\pi}{2n}\right)^4.$$

The above yields

$$\sum_{1\le j\le n/2}\left|b_j^{(n)}-b_j\right|\le\sum_{1\le j\le n/2}\frac{2}{7}\left(\frac{j\pi}{2n}\right)^4 4n^2t\left(1-\left(\frac{j\pi}{2n}\right)^2+\frac{1}{2}\left(\frac{j\pi}{2n}\right)^4\right)^{4n^2t-1}.$$

Now we need a lemma.

LEMMA.

$$\sum_{1\le j\le n/2}\left(\frac{j\pi}{2n}\right)^4\left(1-\left(\frac{j\pi}{2n}\right)^2+\frac{1}{2}\left(\frac{j\pi}{2n}\right)^4\right)^{k-1}\frac{\pi}{2n}\le\frac{9}{k^2}\cdot\frac{\pi}{2n}+\frac{9}{8}\sqrt{\frac{3}{2}}\,\frac{1}{k^2}.$$

Proof. We have

$$\sum_{1\le j\le n/2}\left(\frac{j\pi}{2n}\right)^4\left(1-\left(\frac{j\pi}{2n}\right)^2+\frac{1}{2}\frac{j\pi}{2n}^4\right)^{k-1}\frac{\pi}{2n}$$

$$\le\sum_{1\le j\le n/2}\left(\frac{j\pi}{2n}\right)^4\left(1-\frac{2}{3}\left(\frac{j\pi}{2n}\right)^2\right)^{k-1}\frac{\pi}{2n}.$$

This last sum is a Riemann sum for $\int_0^{\pi/4}\theta^4(1-\frac{2}{3}\theta^2)^{k-1}\,d\theta$. Note that $\theta^4(1-\frac{2}{3}\theta^2)^{k-1}$ rises and falls once and is bounded by $9/k^2$ on $[0,\frac{\pi}{4}]$. Thus the Riemann sum is a lower sum except possibly for one term bounded by $(9/k^2)\pi/2n$. Thus

$$\sum_{1\le j\le n/2}\left(\frac{j\pi}{2n}\right)^4\left(1-\frac{2}{3}\left(\frac{j\pi}{2n}\right)^2\right)^{k-1}\frac{\pi}{2n}\le\frac{9}{k^2}\frac{\pi}{2n}+\int_0^{\pi/4}\theta^4\left(1-\frac{2}{3}\theta^2\right)^{k-1}d\theta.$$

Also

$$\int_0^{\pi/4}\theta^4\left(1-\frac{2}{3}\theta^2\right)^{k-1}d\theta$$

$$=\int_{1-(2/3)(\pi/4)^2}^1\frac{9}{8}\sqrt{\frac{3}{2}}\,(1-u)^{3/2}u^{k-1}\,du\qquad\left(u=1-\frac{2}{3}\theta^2\right)$$

$$\le\frac{9}{8}\sqrt{\frac{3}{2}}\int_0^1(1-u)u^{k-1}\,du=\frac{9}{8}\sqrt{\frac{3}{2}}\,\frac{1}{k(k+1)}$$

$$\le\frac{9}{8}\sqrt{\frac{3}{2}}\,\frac{1}{k^2}.$$

This proves the lemma. □

Continuing, we have

$$\sum_{1\le j\le n/2}\left|b_j^{(n)}-b_j\right|\le\frac{2}{7}4n^2t\,\frac{2n}{\pi}\sum_{1\le j\le n/2}\left(\frac{j\pi}{2n}\right)^2\left(1-\left(\frac{j\pi}{2n}\right)^2+\frac{1}{2}\left(\frac{j\pi}{2n}\right)^4\right)^{4n^2t-1}\frac{\pi}{2n}$$

$$\le\frac{2\cdot4n^2t}{7}\,\frac{2n}{\pi}\left(\frac{9}{16n^4t^2}\cdot\frac{\pi}{2n}+\frac{9}{8}\sqrt{\frac{3}{2}}\,\frac{1}{16n^4t^2}\right)$$

$$\le\frac{9}{7}\left(\frac{1}{\sqrt{10}}+\frac{\sqrt{15}}{4\pi}\right)\frac{1}{n\sqrt{t}}\le.8031\,\frac{1}{n\sqrt{t}}.$$

This proves the seventh estimate.

For the eighth and final estimate we have

$$\|c^{(n)}-c\|\le\left|\left(\frac{2}{1+\frac{1}{n}}\right)^{1/2}-2^{1/2}\right|\le\frac{1}{\sqrt{2}}\,\frac{1}{n}.$$

3. Calculation of the error bound for the final result. Recall from the introduction the distribution functions

$$F_{x,t}^{(m)}(z), \qquad x,z\in\left\{0,\frac{1}{m},\cdots,\frac{k}{m},\cdots,1\right\},$$
$$t\in\left\{0,\frac{1}{2m^2},\frac{2}{2m^2},\cdots,\frac{N}{2m^2},\cdots\right\}.$$

We have $G_{x,t}^{(m/2)}(z)=F_{x,t}^{(m)}(z)$, if m is even, $N=2m^2t$ is even, x is an even multiple of $\frac{1}{m}$. Thus

$$\left|F_{x,t}(z)-F_{x,t}^{(m)}(z)\right|\leq 1.8\,\frac{2}{m\sqrt{t}}=3.6\,\frac{1}{m\sqrt{t}}=3.6\,\frac{\sqrt{2}}{\sqrt{N}}.$$

REFERENCES

[1] K. L. CHUNG, *A Course in Probability Theory*, Harcourt Brace & World, New York, 1968.

[2] S. N. ETHIER AND F. M. NORMAN, *Error estimate for the diffusion approximation of the Wright-Fisher model*, Proc. Nat. Acad. Sci. U.S.A., 74 (1977), pp. 5096–5098.

[3] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. I, John Wiley, New York, 1968.

[4] P. D. LAX, *A stability theorem for solutions of abstract differential equations, and its application to the study of the local behavior of solutions of elliptic equations*, Comm. Pure Appl. Math., 9 (1956), pp. 747–766.

[5] A. G. MACKIE, *Boundary Value Problems*, Oliver & Boyd, Edinburgh, 1965.

[6] H. P. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.

[7] F. L. SPITZER, *Principles of Random Walks*, Springer-Verlag, New York, 1976.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0011 $01.25/0

THE REFORMULATION OF AN INFINITE SUM VIA SEMIINTEGRATION*

KEITH B. OLDHAM†

Abstract. Using the operations of the fractional calculus, the chemically important sum $-\Sigma(-1)^j j^{1/2}\exp(jx)$ is proved equivalent to $\Sigma\beta^{-3}[\pi(\beta-x)/2]^{1/2}[\beta+2x]$ where $\beta=[(2j-1)^2\pi^2+x^2]^{1/2}$ and where, in each summation, j runs from 1 to ∞. Seven other sums of exponential functions are similarly reformulated, as well as eight summations of sine or cosine functions, of which $\Sigma j^{-1/2}\sin(j\pi x)$ is representative. The utility of these reformulations is demonstrated.

The fractional calculus has been shown to be a useful tool in the evaluation of definite integrals, the summation of series, the solution of differential equations and in other areas of mathematics [1], [2]. In this short article, the fractional operations of semiintegration and semidifferentiation will be used to convert an important infinite sum into a more useful expression and to perform similar conversions on related summations.

Weyl semioperations. The expressions

$$\frac{d^{-1/2}f(x)}{[d(x-a)]^{-1/2}}=\pi^{-1/2}\int_a^x[x-y]^{-1/2}f(y)\,dy \tag{1}$$

and

$$\frac{d^{1/2}f(x)}{[d(x-a)]^{1/2}}=\pi^{-1/2}\frac{d}{dx}\int_a^x[x-y]^{-1/2}f(y)\,dy \tag{2}$$

establish the notation for, and one definition of, the generalized semiintegration and semidifferentiation operations on any function f of an independent variable x [1]. The parameter a plays the role of a lower limit. Here we are primarily concerned with semioperations in which $a=-\infty$; such operations are known as Weyl semiintegration and Weyl semidifferentiation [3]. Bateman [4, p. 201] used a slightly different definition of a Weyl fractional integral.

Rather comprehensive tables of semiintegrals and semiderivatives exist [1], but these refer to the $a=0$ instance of definitions (1) and (2) and there is no generally applicable method of translating the lower limit in the fractional calculus. Accordingly, Table 1 is presented to include all the instances required in this note. Entries #1 through #5 are elementary, but the semiintegral entry #6 was established by specializing the general result

$$\frac{d^{-1/2}}{[d(x-a)]^{-1/2}}\frac{x}{b^2+x^2}=\frac{1}{|2\pi x|^{1/2}S}\left((S\pm1)^{1/2}\operatorname{arctanh}\frac{R(2S\pm2)^{1/2}}{S+R^2}-(S\mp1)^{1/2}\operatorname{Arctan}\frac{R(2S\mp2)^{1/2}}{S-R^2}\right) \tag{3}$$

* Received by the editors June 5, 1979, and in revised form November 5, 1981.
† Chemistry Department, Trent University, Peterborough, Ontario, Canada K9J 7B8.

to $a=-\infty$. The semiderivative then follows by differentiation of the semiintegral entry. The formulas tabulated as #7 may be derived similarly. The derivation of (3), in which R and S are abbreviations for $|1-(a/x)|^{1/2}$ and $\{1+(b/x)^2\}^{1/2}$ respectively and in which the upper or lower signs apply according as x is positive or negative, was facilitated by a change of the integration variable in definition (1) to $|1-(y/x)|^{1/2}$ and partial fractioning of the resulting rational integrand.

Series reformulation. The function $\chi(x)$ occurs in electrochemistry, where it describes the dependence of electric current on cell voltage under certain electrolysis conditions [5], [6], [7]. The $\chi(x)$ versus x graph is an asymmetrically peaked curve: this and similar curves are used by electrochemists to study electrode reactions and to perform chemical analyses. Table 2 lists values of $\pi^{1/2}\chi(x)$. Reinmuth [8] showed that the function is described by the summation

$$\pi^{1/2}\chi(x)=-\sum_{j=1}^{\infty}(-1)^j j^{1/2}\exp(jx), \qquad x<0 \tag{4}$$

for negative x and this provides a convenient means for the calculation of numerical values of $\chi(x)$ in that range of argument.

However, the chemically interesting portion of the χ curve lies in the region of its peak, which as Table 2 shows corresponds to positive x values, and therefore less attractive computational techniques have had to be employed to determine $\chi(x)$ for $x>0$. These techniques include the numerical solution of the corresponding integral equation [9] and the numerical quadrature of a related definite integral [10], [11]. Especially when experimental data are to be processed by a computer, there are marked advantages in an analytical expression over solutions in tabular form [12]. Accordingly, one purpose of this study is to analytically continue formula (4) to the range $x\geq 0$.

Using entry #1 of Table 1, function (4) may be Weyl semiintegrated [1], yielding the closed form expression

$$\pi^{1/2}\frac{d^{-1/2}\chi(x)}{[d(x+\infty)]^{-1/2}}=-\sum_{j=1}^{\infty}(-1)^j\exp(jx)=\frac{1}{2}+\frac{1}{2}\tanh\frac{x}{2}, \qquad x<0. \tag{5}$$

By differentiation with respect to x of the infinite product expression [13, p. 85] for the hyperbolic cosine of $x/2$, one can establish that

$$\tanh\frac{x}{2}=4\sum_{k=1}^{\infty}\frac{x}{b_k^2+x^2}, \qquad b_k=(2k-1)\pi \tag{6}$$

and substitute this expression into (5). At this point semidifferentiation is applied to both sides of the relationship. Because semidifferentiation of a semiintegral regenerates the original function [1], [14], [15], the left side becomes $\pi^{1/2}\chi(x)$. The tabulated semiderivatives allow Weyl semidifferentiation of the right side of the equation and lead to

(7)

$$\pi^{1/2}\chi(x)=\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}\beta^{-3}[\beta-x]^{1/2}[\beta+2x], \qquad \beta=\left[(2k-1)^2\pi^2+x^2\right]^{1/2}, \qquad x<0$$

as the final expression for $\chi(x)$.

Because (4) and (7) are equivalent for negative x, and summation (7) converges for all values of x, the latter constitutes an analytic continuation of $\pi^{1/2}\chi(x)$. For some x

values, the convergence of summation (7) is rather slow. To hasten the process, advantage may be taken of the expressibility of the summand as

$$\beta^{-3}[\beta-x]^{1/2}[\beta+2x]=b_k^{-3/2}+\frac{3}{2}xb_k^{-5/2}-\frac{15}{8}x^2b_k^{-7/2}+O\left(x^3b_k^{-9/2}\right)$$

to rewrite (7) in the form

$$\pi^{1/2}\chi(x)=\pi^{1/2}\chi(0)+\pi^{1/2}\chi'(0)x+\frac{1}{2}\pi^{1/2}\chi''(0)x^2 + \left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}\frac{[\beta-x]^{1/2}[\beta+2x]}{\beta^3}-\frac{b_k^2+\frac{3}{2}b_kx-\frac{15}{8}x^2}{b_k^{7/2}}, \tag{8}$$

where

$$\pi^{1/2}\chi(0)=\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}b_k^{-3/2}=\frac{[8^{1/2}-1]\zeta(\frac{3}{2})}{4\pi}=0.380104813,$$

$$\pi^{1/2}\chi'(0)=\frac{3}{2}\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}b_k^{-5/2}=\frac{3\left[(32)^{1/2}-1\right]\zeta(\frac{5}{2})}{16\pi^2}=0.118680871$$

and

$$\frac{\pi^{1/2}}{2}\chi''(0)=\frac{-15}{8}\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}b_k^{-7/2}=\frac{-15\left[(128)^{1/2}-1\right]\zeta(\frac{7}{2})}{128\pi^3}=-0.043920560.$$

Formula (8) proves to be a highly efficient method for computing $\pi^{1/2}\chi(x)$ and was, in fact, employed in the construction of Table 2 with the aid of a pocket calculator. The data in the tabulation are consistent with, but more precise than, previously published values of the function [9].

Equations (4) and (7) are equivalent series representations of the $\pi^{1/2}\chi(x)$ function. That the series

$$\pi^{1/2}\chi(x)=\frac{1}{(\pi x)^{1/2}}\left\{1+\frac{\pi^2}{8x^2}+\frac{49\pi^4}{384x^4}+\frac{341\pi^6}{1024x^6}+O(x^{-8})\right\} \tag{9}$$

provides a third alternative, valid for large positive x, will be established later in this article.

Similar relationships. A number of series reformulations are possible via hyperbolic functions other than the tangent. By methods similar to that described for the $\chi(x)$ function, one can establish that

$$\sum_{j=1}^{\infty}j^{1/2}\exp(-jx)=\frac{\pi^{1/2}}{2x^{3/2}}-\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}\frac{[\alpha+x]^{1/2}[\alpha-2x]}{\alpha^3}, \tag{10}$$

$$\sum_{j=1}^{\infty}(-1)^jj^{1/2}\exp(-jx)=-\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}\frac{[\beta+x]^{1/2}[\beta-2x]}{\beta^3}, \tag{11}$$

$$\sum_{j=1}^{\infty}(2j-1)^{1/2}\exp(x-2jx)=\frac{\pi^{1/2}}{4x^{3/2}}-\left(\frac{\pi}{8}\right)^{1/2}\sum_{k=1}^{\infty}(-1)^k\frac{[\gamma+x]^{1/2}[\gamma-2x]}{\gamma^3}, \tag{12}$$

and

$$(13)\quad \sum_{j=1}^{\infty}(-1)^j(2j-1)^{1/2}\exp(x-2jx)=\left(\frac{\pi}{8}\right)^{1/2}\sum_{k=1}^{\infty}(-1)^k\frac{[\delta-x]^{1/2}[\delta+2x]}{\delta^3},$$

where $\alpha=[4k^2\pi^2+x^2]^{1/2}$, $\beta=[(2k-1)^2\pi^2+x^2]^{1/2}$, $\gamma=[k^2\pi^2+x^2]^{1/2}$ and $\delta=[(k-\frac{1}{2})^2\pi^2+x^2]^{1/2}$. Ranges of convergence vary among the eight summations in these equations, but all are valid for $x>0$.

Area beneath the $\chi(x)$ curve. Electrochemically, the area beneath the $\pi^{1/2}\chi(x)$ curve corresponds to the quantity of electricity that has passed through the electrolysis cell. To analytically continue

$$(14)\quad \pi^{1/2}\int_{-\infty}^{x}\chi(y)\,dy=-\sum_{j=1}^{\infty}(-1)^j\frac{\exp(jx)}{j^{1/2}},\qquad x<0$$

is therefore of interest. This summation is only marginally more tractible than that in (4). It is possible to reformulate the series in (14) by semiintegration of (5) and (6). However, a simpler procedure will be followed.

Combination of (4) and (7) to

$$-\sum_{j=1}^{\infty}(-1)^j j^{1/2}\exp(jx)=\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}\frac{[\beta-x]^{1/2}[\beta+2x]}{\beta^3},\qquad x<0,$$

followed by integration with an upper limit of zero, leads to

(15)

$$\sum_{j=1}^{\infty}(-1)^j\frac{1}{j^{1/2}}-\sum_{j=1}^{\infty}(-1)^j\frac{\exp(jx)}{j^{1/2}}=(2\pi)^{1/2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^{1/2}\pi^{1/2}}-\frac{[\beta-x]^{1/2}}{\beta}.$$

The equation

$$(16)\quad \pi^{1/2}\int_{-\infty}^{x}\chi(y)\,dy=-B+(2\pi)^{1/2}\sum_{k=1}^{\infty}\frac{1}{b_k^{1/2}}-\frac{\left[\left(b_k^2+x^2\right)^{1/2}-x\right]^{1/2}}{\left(b_k^2+x^2\right)^{1/2}}$$

in which b_k is again used as an abbreviation for $(2k-1)\pi$ and

$$(17)\quad B=\sum_{j=1}^{\infty}(-1)^j\frac{1}{j^{1/2}}=[2^{1/2}-1]\zeta\left(\frac{1}{2}\right)=-0.60489864$$

then results from the union of (14) and (15). Equation (16) provides an expression for the area under the $\pi^{1/2}\chi(x)$ curve that converges for all values of x.

There is interest in determining the form of the integral of $\pi^{1/2}\chi(x)$ for large x. In this circumstance use may be made of an Euler–Maclaurin transformation [13, p. 806] of the summation in (16), as follows

$$\pi^{1/2}\int_{-\infty}^{x}\chi(y)\,dy=-B+(2\pi)^{1/2}\left\{\int_1^{\infty}F(k)\,dk+\frac{1}{2}F(1)-\frac{1}{12}F'(1)+\frac{1}{720}F'''(1)-\cdots\right\},$$

where $F(k)$ represents the summand in (16) and primes denote differentiation with respect to k. Thereby one may show that

$$\pi^{1/2}\int_{-\infty}^{x}\chi(y)\,dy=\text{constant}+2\left(\frac{x}{\pi}\right)^{1/2}\left\{1-\frac{\pi^2}{24x^2}-\frac{7\pi^4}{384x^4}-\frac{31\pi^6}{1024x^6}+O(x^{-8})\right\}.$$

It is the differentiation of this result that produces (9).

Similar integrated relationships. Equation (16) is essentially the integral of (11) using an integration limit of zero. Though differing limits are advantageous, all of (10) through (13) can be treated similarly. The results are

$$(18)\quad \sum_{j=1}^{\infty}\frac{\exp(-jx)}{j^{1/2}}=A+\left(\frac{\pi}{x}\right)^{1/2}+(2\pi)^{1/2}\sum_{k=1}^{\infty}\frac{[\alpha+x]^{1/2}}{\alpha}-\frac{\left[(4k^2+1)^{1/2}+1\right]^{1/2}}{\pi^{1/2}[4k^2+1]^{1/2}},$$

$$(19)\quad \sum_{j=1}^{\infty}(-1)^j\frac{\exp(-jx)}{j^{1/2}}=B+(2\pi)^{1/2}\sum_{k=1}^{\infty}\frac{[\beta+x]^{1/2}}{\beta}-\frac{1}{[2k-1]^{1/2}\pi^{1/2}},$$

$$(20)\quad \sum_{j=1}^{\infty}\frac{\exp(x-2jx)}{(2j-1)^{1/2}}=\frac{C}{2^{1/2}}+\frac{1}{2}\left(\frac{\pi}{x}\right)^{1/2}+\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}(-1)^k\left\{\frac{[\gamma+x]^{1/2}}{\gamma}-\frac{\left[(k^2+1)^{1/2}+1\right]^{1/2}}{\pi^{1/2}[k^2+1]^{1/2}}\right\},$$

$$(21)\quad \sum_{j=1}^{\infty}(-1)^j\frac{\exp(x-2jx)}{(2j-1)^{1/2}}=\left(\frac{\pi}{2}\right)^{1/2}\sum_{k=1}^{\infty}(-1)^k\frac{[\delta-x]^{1/2}}{\delta},$$

where $A=-1+\Sigma j^{-1/2}\exp(-j\pi)=-0.9554172$, B is given by (17) and $C=-2^{-1/2}+\Sigma(j-\frac{1}{2})^{-1/2}\exp(\pi-2j\pi)=-0.64592709$.

Using Euler's formula, the exponential summations in (10) through (13) may be replaced by trigonometric summations; however they generally diverge. The same is not true of transformations of (18) through (21). A catalog of the resulting trigonometric sums is

$$\sum_{j=1}^{\infty}\frac{\sin(j\pi x)}{j^{1/2}}=\frac{1}{2^{1/2}x^{1/2}}-\sum_{k=1}^{\infty}\frac{[2k-\kappa]^{1/2}}{\kappa},$$

$$\sum_{j=1}^{\infty}\frac{\cos(j\pi x)}{j^{1/2}}=A+\frac{1}{2^{1/2}x^{1/2}}+\sum_{k=1}^{\infty}\frac{[2k+\kappa]^{1/2}}{\kappa}-\frac{\left[(4k^2+1)^{1/2}+1\right]^{1/2}}{\left[2k^2+\frac{1}{2}\right]^{1/2}},$$

$$\sum_{j=1}^{\infty}(-1)^j\frac{\sin(j\pi x)}{j^{1/2}}=-\sum_{k=1}^{\infty}\frac{[2k-1-\lambda]^{1/2}}{\lambda},$$

$$\sum_{j=1}^{\infty}(-1)^j\frac{\cos(j\pi x)}{j^{1/2}}=B+\sum_{k=1}^{\infty}\frac{[2k-1+\lambda]^{1/2}}{\lambda}-\frac{1}{\left[k-\frac{1}{2}\right]^{1/2}},$$

$$\sum_{j=1}^{\infty}\frac{\sin\left[(j-\frac{1}{2})\pi x\right]}{(j-\frac{1}{2})^{1/2}}=\frac{1}{2^{1/2}x^{1/2}}-\sum_{k=1}^{\infty}(-1)^k\frac{[2k-\kappa]^{1/2}}{\kappa},$$

$$\sum_{j=1}^{\infty}\frac{\cos[(j-\frac{1}{2})\pi x]}{(j-\frac{1}{2})^{1/2}}=C+\frac{1}{2^{1/2}x^{1/2}}+\sum_{k=1}^{\infty}(-1)^k\left\{\frac{[2k+\kappa]^{1/2}}{\kappa}-\frac{\left[(k^2+1)^{1/2}+1\right]^{1/2}}{[k^2+1]^{1/2}}\right\},$$

$$\sum_{j=1}^{\infty}(-1)^j\frac{\sin[(j-\frac{1}{2})\pi x]}{(j-\frac{1}{2})^{1/2}}=\sum_{k=1}^{\infty}(-1)^k\frac{[2k-1-\lambda]^{1/2}}{\lambda},$$

$$\sum_{j=1}^{\infty}(-1)^j\frac{\cos[(j-\frac{1}{2})\pi x]}{(j-\frac{1}{2})^{1/2}}=\sum_{k=1}^{\infty}(-1)^k\frac{[2k-1+\lambda]^{1/2}}{\lambda},$$

where $\kappa=[4k^2-x^2]^{1/2}$ and $\lambda=[(2k-1)^2-x^2]^{1/2}$. Again, the range of validity of these equations is various, but all hold for $0<x<1$.

Power series useful for small x can be constructed from these eight equations; the first, for example, yields

$$\sum_{j=1}^{\infty}\frac{\sin(j\pi x)}{j^{1/2}}=\frac{1}{2^{1/2}x^{1/2}}-\frac{\zeta(\frac{3}{2})}{4}x-\frac{5\zeta(\frac{7}{2})}{128}x^3-\frac{63\zeta(\frac{11}{2})}{8192}x^5+O(x^7).$$

Moreover these equations can generate novel numerical series; thus the first can be reduced to

$$\sum_{k=1}^{\infty}\frac{\left[k-(k^2-\frac{1}{4})^{1/2}\right]^{1/2}}{\left[k^2-\frac{1}{4}\right]^{1/2}}=1$$

after x is set to unity.

TABLE 1

The Weyl semiintegrals and Weyl semiderivatives of seven functions

entry	$f(x)$	$\dfrac{d^{-1/2}f(x)}{[d(x+\infty)]^{-1/2}}$	$\dfrac{d^{1/2}f(x)}{[d(x+\infty)]^{1/2}}$
#1	$\exp(bx),\ b>0$	$b^{-1/2}\exp(bx)$	$b^{1/2}\exp(bx)$
#2	$\sin(bx)$	$b^{-1/2}\sin(bx-\pi/4)$	$b^{1/2}\sin(bx+\pi/4)$
#3	$\cos(bx)$	$b^{-1/2}\cos(bx-\pi/4)$	$b^{1/2}\cos(bx+\pi/4)$
#4	any constant	∞	0
#5	$\dfrac{1}{b-x}$	$\left(\dfrac{\pi}{b-x}\right)^{1/2}$	$\dfrac{\pi^{1/2}}{2(b-x)^{3/2}}$
#6	$\dfrac{x}{b^2+x^2}$	$-\dfrac{\pi^{1/2}\left[(b^2+x^2)^{1/2}-x\right]^{1/2}}{2^{1/2}[b^2+x^2]^{1/2}}$	$\dfrac{\pi^{1/2}\left[(b^2+x^2)^{1/2}-x\right]^{1/2}\left[(b^2+x^2)^{1/2}+2x\right]}{2^{3/2}[b^2+x^2]^{3/2}}$
#7	$\dfrac{b}{b^2+x^2}$	$\dfrac{\pi^{1/2}\left[(b^2+x^2)^{1/2}+x\right]^{1/2}}{2^{1/2}[b^2+x^2]^{1/2}}$	$\dfrac{\pi^{1/2}\left[(b^2+x^2)^{1/2}+x\right]^{1/2}\left[(b^2+x^2)^{1/2}-2x\right]}{2^{3/2}[b^2+x^2]^{3/2}}$

TABLE 2

Values and features of the function $\chi(x)$

x	$\pi^{1/2}\chi(x)$	
$\to -\infty$	$\exp(x)$	limiting expression
−9.0000	0.00012	
−8.0000	0.00034	
−7.0000	0.00091	
−6.0000	0.00247	
−5.0000	0.00667	
−4.0000	0.01785	
−3.0000	0.04648	
−2.0000	0.11314	
−1.0934	0.22315	half peak
−1.0000	0.23681	
−0.7315	0.27719	inflection point
0.0000	0.38010	
+1.0000	0.44572	
1.1090	0.44629	peak
2.0000	0.41815	
2.5950	0.38362	inflection point
3.0000	0.35951	
4.0000	0.30747	
5.0000	0.26886	
6.0000	0.24093	
6.8400	0.22315	half peak
7.0000	0.22020	
8.0000	0.20427	
9.0000	0.19146	
10.0000	0.18093	
$\to \infty$	$(\pi x)^{-1/2}$	limiting expression

REFERENCES

[1] K. B. OLDHAM AND J. SPANIER, *The Fractional Calculus*, Academic Press, New York & London, 1974.

[2] B. ROSS, ed., *Fractional Calculus and Its Applications*, Lecture Notes in Mathematics 457, Springer-Verlag, Berlin, Heidelberg & New York, 1975.

[3] H. WEYL, Vierteljschr. Naturforsch. Gesellsch. Zurich, 62 (1917), pp. 296–302.

[4] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Tables of Integral Transforms*, Vol. II, McGraw-Hill, New York, Toronto and London, 1954.

[5] J. E. B. RANDLES, *Cathode-ray polarography*. II. *Current-voltage curves*, Trans. Faraday Soc., 44 (1948), pp. 327–338.

[6] A. SEVCIK, *Oscillographic polarography with periodic triangular voltage*, Collect. Czech. Chem. Commun., 13 (1948), pp. 349–377.

[7] P. DALRYMPLE-ALFORD, M. GOTO AND K. B. OLDHAM, *Shapes of derivative neopolarograms*, J. Electroanal. Chem, 85 (1977), 1–15.

[8] W. H. REINMUTH, *Theory of diffusion-limited charge-transfer processes in electroanalytical techniques*, Analytical Chemistry, 34 (1962), pp. 1446–1454.

[9] R. S. NICHOLSON AND I. SHAIN, *Theory of stationary electrode polarography. Single scan and cyclic methods applied to reversible, irreversible and kinetic systems*, Analytical Chemistry, 36 (1964), pp. 706–723.

[10] H. MATSUDA AND Y. AYABE, *The theory of cathode-ray polarography of Randles-Sevcik*, Z. Elektrochem., 59 (1955), pp. 494–503.

[11] Y. P. GOKHSTEIN, *General equations for oscillographic polarography; reversible processes in cathode and anode polarization*, Doklady Akad. Nauk SSSR, 126 (1959), pp. 598–601. (In Russian.)

[12] K. B. OLDHAM, *Analytical expressions for the reversible Randles-Sevcik function*, J. Electroanal. Chem., 105 (1979), pp. 373–375.

[13] M. ABRAMOWITZ AND I. STEGUN, eds., *Handbook of Mathematical Functions*, N.B.S. Applied Mathematics Series #55, U.S. Government Printing Office, Washington, DC, 1964.

[14] K. B. OLDHAM AND J. SPANIER, *Fractional Calculus and Its Applications*, Bulletin of the Polytechnic Institute of Jassy, XXIV(XXVII) (1978), pp. 29–34.

[15] K. B. OLDHAM, *Application of the fractional calculus to heat transport and electrochemistry*, in Physicochemical Hydrodynamics, D. B. Spalding, ed., Advance Publications, London, 2 (1977), pp. 1045–1054.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0012 $01.25/0

MEAN GROWTH, H_p SPACES AND SUBHARMONIC FUNCTIONS IN THE UPPER HALF-PLANE*

JAMES D. McCALL†

Abstract. A function f analytic in the upper half-plane is said to be of class H_p $(0<p<\infty)$ if the L_p integrals of $f_y(x)=f(x+iy)$ along lines parallel to the real axis are uniformly bounded. In this paper we give alternate proofs of two results of B. F. Logan [SIAM J. Math. Anal., 10 (1979), pp. 733–740; 741–751] on H_p spaces plus a subharmonic version of his second result. Our method of proof extends his H_p results from $1\le p<\infty$ to $0<p<\infty$. We show that if $f\in H_p$ $(0<p<\infty)$, then

$$\lim_{y\to\infty}\|f_y\|_p=0$$

and

$$|f(x+iy)|\le A_p y^{-1/p}\left\{\int_{-\infty}^{\infty}|f(t)|^p\,dt\right\}^{1/p}$$

with $A_p=(4\pi)^{-1/p}$. The constant A_p is best possible and necessary and sufficient conditions are given for equality.

The method of proof in each case consists in proving it for $p=2$ by means of the one-sided Paley–Wiener theorem and then extending it to $0<p<\infty$ by using the Blaschke factorization for H_p functions.

The subharmonic analogue for the second result shows that for a continuous nonnegative subharmonic function $g(z)$ defined in the upper half-plane, the condition

$$\int_{-\infty}^{\infty}g(x+iy)\,dx\le M<\infty$$

for $y>0$ implies

$$g(x+iy)\le M(\pi y)^{-1}.$$

The constant π^{-1} is best possible.

Key words. H_p spaces, subharmonic functions, Paley–Wiener theorem, Blaschke factorization

1. Introduction.

A function f analytic in the upper half-plane is said to be of class H_p $(0<p<\infty)$ if the integrals

$$\int_{-\infty}^{\infty}|f(x+iy)|^p\,dx$$

are uniformly bounded with respect to $y>0$. Let

$$f_y(x)=f(x+iy)$$

and

$$\|f_y\|_p=\left\{\int_{-\infty}^{\infty}|f(x+iy)|^p\,dx\right\}^{1/p}.$$

The latter is a pth mean of f. Then $f_y(x)$ tends to a boundary function $f(x)$ a.e. as y tends to zero and, in fact, f_y converges to f in $L_p(-\infty,\infty)$. The quantity $\|f\|_p$ is used as the norm of f even if $0<p<1$ [1, pp. 187–192].

We are ready to state our first two results.

THEOREM 1. *For $f\in H_p$ $(0<p<\infty)$,*

$$\lim_{y\to\infty}\|f_y\|_p=0. \tag{1}$$

*Received by the editors May 28, 1982.

†Department of Mathematics, Texas A & I University, Kingsville, Texas 78363.

THEOREM 2. *For* $f \in H_p$ $(0<p<\infty)$

$$|f(x+iy)|<A_p y^{-1/p}\|f\|_p \tag{2}$$

with $A_p=(4\pi)^{-1/p}$. *The constant* A_p *is best possible and equality holds for* $x=0$ *and* $y=b>0$ *if and only if*

$$f(z)=\frac{C}{(z+ib)^{2/p}}. \tag{3}$$

These were first proved for $1<p<\infty$ by B. F. Logan [5], [6]. In fact he proved the first theorem for a more general class of functions consisting of convolution transforms. However, with respect to H_p classes his proof will not extend to $0<p<1$ because it depends on the Poisson representation which is valid only for $1\leq p<\infty$. Similarly the proof for his second result will not extend because it starts with the Cauchy representation.

Our extensions are accomplished by means of the one-sided Paley–Wiener theorem for the case $p=2$ and then applying the Blaschke factorization for H_p functions to extend them to $0<p<\infty$. Using the Blaschke factorization in this manner is a standard H_p technique [2, pp. 39–41].

The third result is suggested by two things, first, by Theorem 2 and the fact that $g(z)=|f(z)|^p$ is subharmonic, and secondly, by the following lemma due to V. I. Krylov [1, p. 188], [4, pp. 38–40].

LEMMA. *Let* g *be a continuous nonnegative function subharmonic in the upper half-plane satisfying*

$$\int_{-\infty}^{\infty} g(x+iy)\,dx\leq M<\infty \tag{4}$$

for $y>0$. *Then*

$$g(x+iy)\leq M\left(\frac{4}{3\pi y}\right) \tag{5}$$

We will later use this lemma to help prove Theorem 3.

Now using the Poisson kernel as an indicator of the best possible constant, we arrive at the following.

THEOREM 3. *Let* g *be a subharmonic function satisfying the conditions of the lemma. Then*

$$g(x+iy)\leq M(\pi y)^{-1} \tag{6}$$

and the constant π^{-1} *is best possible.*

2. Proof of Theorem 1. Since the Blaschke factorization is used in this and the next section, we give what is needed here. This factorization means for each f in H_p $(0<p<\infty)$, not identically zero, there exist functions b and g such that $f=bg$ and

(i) $g\in H_p$, $\|f\|_p=\|g\|_p$, g does not vanish;

(ii) $|b(z)|\leq 1$ and $|b(x)|=1$ a.e.

The function $b(z)$ is the Blaschke product for the zeros of f. Its form may be found in [1, p. 191].

We now proceed with the proof. According to the one-sided Paley–Wiener theorem for each f in H_2 there is a F in $L_2(0,\infty)$ such that

$$f(z)=\int_0^\infty F(t)e^{izt}\,dt.$$

By Plancherel's theorem

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}|f(x+iy)|^2\,dx=\int_0^\infty|F(t)|^2e^{-2ty}\,dt. \tag{7}$$

Clearly the integrand on the right tends to zero a.e. as y approaches infinity and is dominated by $|F|^2$, so by dominated convergence, the right-hand side of (7) converges to zero as y approaches infinity. This gives the result for $p=2$.

Next let f be a function in H_p $(0<p<\infty)$ that does not vanish. Then $f^{p/2}$ belongs to H_2 and $\|f_y^{p/2}\|_2=\|f_y\|_p^{p/2}$. The case $p=2$ applied to the first mean gives the result.

Finally, let f belong to H_p with $f=bg$, the Blaschke factorization. We have $|f(z)|\le|g(z)|$ where g does not vanish and belongs to H_p. This means that $\|f_y\|_p\le\|g_y\|_p$ with $\|g_y\|_p$ tending to zero as y approaches infinity.

3. Proof of Theorem 2. We again assume $p=2$ and use the Paley–Wiener theorem to write

$$f(x+iy)=\int_0^\infty F(t)e^{i(x+iy)t}\,dt$$

with F in $L_2(0,\infty)$. Applying the Schwarz inequality and Plancherel's theorem in succession, we have

$$|f(x+iy)|\le\|f\|_2(4\pi y)^{-1/2}$$

with equality for $x=0$ and $y=b>0$ if and only if

$$F(t)=Be^{-bt}\quad\text{a.e. on }(0,\infty).$$

This means

$$f(z)=B\int_0^\infty e^{-bt}e^{izt}\,dt=\frac{iB}{(z+ib)}.$$

The equality condition implies $(4\pi)^{-1/2}$ is best possible. Next suppose f belongs to H_p and does not vanish. Then $f^{p/2}$ is in H_2 and

$$|f(x+iy)|^{p/2}\le\|f^{p/2}\|_2(4\pi y)^{-1/2}.$$

Since $\|f_y^{p/2}\|_2=\|f_y\|_p^{p/2}$, the result follows. Equality occurs for $x=0$ and $y=b>0$ if and only if

$$f^{p/2}(z)=\frac{C}{z+ib}.$$

This is certainly equivalent to the condition in the theorem. Because of this condition $(4\pi)^{-1/p}$ has to be the best possible.

Finally suppose $f\in H_p$. Then $f=bg$ where $\|f\|_p=\|g\|_p$, g does not vanish and $|b(z)|\le1$. Thus

$$|f(x+iy)|\le|g(x+iy)|\le\|g\|_p(4\pi y)^{-1/p}=\|f\|_p(4\pi y)^{-1/p},$$

with equality if and only if $b\equiv 1$ and

$$g(z)=\frac{C}{(z+ib)^{2/p}}.$$

The constant is best possible.

4. Proof of Theorem 3. Using an argument of Logan's [6, pp. 742–43], we first reduce the theorem to showing

$$g(i)\leq M\pi^{-1}. \tag{8}$$

First, assume that (4) implies

$$g(iy)\leq M(\pi y)^{-1} \tag{9}$$

for $y>0$. Next let $h(x+iy)=g(x+a+iy)$ $(-\infty<a<\infty)$. A linear change of variable shows (4) holds for h. Hence (9) holds for h or

$$g(a+iy)\leq M(\pi y)^{-1}$$

for $y>0$.

Now assume (4) implies (8) and let $h(x+iy)=g(bx+iby)$. The change of variable $\bar{x}=bx$ in $\int_{-\infty}^{\infty} g(bx+iby)\,dx$ shows

$$\int_{-\infty}^{\infty} h(x+iy)\,dx\leq Mb^{-1}<\infty.$$

Thus (4) holds for h with the constant equal to Mb^{-1}. This means

$$h(i)\leq M(\pi b)^{-1} \quad \text{or} \quad g(ib)\leq M(\pi b)^{-1}.$$

Thus it remains only to show (8). Map the upper half-plane onto the unit disk by

$$w=\frac{z-i}{z+i}, \qquad z=\frac{i(1+w)}{1-w},$$

and note that $y=b$ is mapped into the circle C_b with center $b(1+b)^{-1}$ and radius $R=(1+b)^{-1}$. Let

$$h(w)=g\left(\frac{i(1+w)}{1-w}\right),$$

which is subharmonic in $|w|<1$ [3, p. 12]. Also let Γ_b be a circle with center $b(1+b)^{-1}$ and radius $\rho<R$. By the local submean value property

$$h\left(\frac{b}{1+b}\right)\leq\frac{1}{2\pi\rho}\int_{\Gamma_b} h(w)\,|dw|.$$

Since subharmonic means are nondecreasing [1, p. 9] and h is bounded on and inside by C_b by Krylov's lemma, an application of the bounded convergence theorem gives

$$h\left(\frac{b}{1+b}\right)\leq\frac{1}{2\pi R}\int_{C_b} h(w)\,|dw|$$

when we let $\rho \to R$. Using $R=(1+b)^{-1}$, $|1-w|\leq 2$, and the change of variable from w to z, we have

$$h\left(\frac{b}{1+b}\right)\leq \frac{1+b}{\pi}\int_{-\infty}^{\infty} g(x+ib)\,dx \leq M\left(\frac{1+b}{\pi}\right).$$

Letting $b\to 0$, we get $h(0)\leq M\pi^{-1}$ or $g(i)\leq M\pi^{-1}$.

We must still show π^{-1} is the best possible. Let

$$P_C(x+iy)=\frac{1}{\pi}\frac{Cy}{x^2+y^2},$$

a multiple of the Poisson kernel. This is harmonic and hence subharmonic. Further, $\int_{-\infty}^{\infty} P_C(x+iy)\,dx = C$ for $y>0$. Taking $x=0$ and $y=b>0$, we find

$$P_C(ib)=C(\pi b)^{-1}.$$

Remark. The argument for $g(i)\leq M\pi^{-1}$ is an adaptation of the proof of Krylov's second lemma [1, p. 189], [4, pp. 40–41]. Here it is proved that (4) implies g has a harmonic majorant.

REFERENCES

[1] P. L. DUREN, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[2] C. L. FEFFERMAN, *Harmonic analysis and H^p spaces*, Studies in Harmonic Analysis, MAA Studies in Mathematics 13, J. M. Ash, ed., Mathematical Association of America, Washington, D C, 1976, pp. 38–75.

[3] W. H. J. FUCHS, *Topics in the Theory of Functions of One Complex Variable*, Van Nostrand, Princeton, N J, 1967.

[4] V. I. KRYLOV, *On functions regular in a half-plane*, Mat. Sb., 6 (48) (1939), pp. 95–113; Amer. Math. Soc. Transl., (2) 32 (1963), pp. 37–81.

[5] B. F. LOGAN, *Limits in L_p of convolution transforms with kernels $aK(at)$, $a\to 0$*, this Journal, 9 (1981), pp. 733–740.

[6] _____, *Inequalities and minimum norm kernels for the Hardy class H_p*, this Journal, 10 (1979), pp. 741–751.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0013 $01.25/0

UNE TRANSFORMATION DE LAPLACE–JACOBI*

MICHEL MIZONY†

Abstract. We define an integral transformation on $\mathbb{R}_+$, starting from Jacobi functions of the second type. This transformation is a generalization of the usual Laplace transform. The reciprocal transformation is defined using functions which are closely related to Jacobi functions of the first and second types. Finally, for the particular case of the $SO_0(1,n)$ groups, we give an interpretation of the Jacobi function of the second type as a mean of a "hyperbolic Poisson kernel."

Présentation. A chaque couple (α,β) de nombres complexes est associée une mesure $\omega_{\alpha,\beta}(t)\,dt$ sur $\mathbb{R}_+$ et à cette mesure un Laplacien: $\Delta_{\alpha,\beta}=(1/\omega_{\alpha,\beta}(t))\frac{d}{dt}(\omega_{\alpha,\beta}(t)\frac{d}{dt})$. Les deux fonctions de Jacobi de 1-ère et de 2-ème espèce $t\to\varphi_{\alpha,\beta}(\lambda,t)$ et $t\to\Phi_{\alpha,\beta}(\lambda,t)$ sont fonctions propres (linéairement indépendantes) de l'opérateur $\Delta_{\alpha,\beta}$ pour la valeur propre $\lambda^2+(\alpha+\beta+1)^2$, et ceci pour chaque $\lambda\in\mathbb{C}$.

La transformation de Jacobi (ou de Fourier–Jacobi), définie pour $f\in L^1(\mathbb{R}_+,\omega_{\alpha,\beta}(t)\,dt)$ par

$$\mathcal{F}_{\alpha,\beta}(f)(\lambda)=\frac{2^{2(\alpha+\beta+1)}\sqrt{2}}{\Gamma(\alpha+1)}\int_0^\infty f(t)\varphi_{\alpha,\beta}(\lambda,t)\omega_{\alpha,\beta}(t)\,dt,$$

a été étudiée par différents auteurs et se réduit à la transformation de Fourier–sphérique sur les groupes de Lie semi-simples non compacts de rang réel 1, pour certaines valeurs demi-entières des paramètres α et β. Nous retiendrons pour notre exposé les méthodes analytiques de démonstrations établies par T. Koornwinder [9].

Par ailleurs, le physicien G. A. Viano [14] souligne l'importance, dans le cas $\beta=-\frac{1}{2}$ $\alpha=0$, d'une transformation intégrale associée à la fonction propre $\Phi_{0,-1/2}(\lambda,t)=Q_{-1/2+i\lambda}(\operatorname{ch} t)$ (c'est la fonction de Legendre de 2-ème espèce).

Cette transformation intervient dans l'étude de problèmes d'amplitude de dispersion en théorie des interactions fortes. G. A. Viano rattache cette transformation à la géométrie de l'espace Riemannien symétrique $SU(1,1)/SO(2)$, et lui donne un statut de transformation de Laplace.

Nous allons dans cet article étudier systématiquement pour $\alpha,\beta\in\mathbb{C}$ la transformation $\mathcal{L}_{\alpha,\beta}$ définie par exemple pour f continue à support compact dans $\mathbb{R}_+^*$ par

$$\mathcal{L}_{\alpha,\beta}(f)(\lambda)=\int_0^\infty f(t)\big(\Phi_{\alpha,\beta}(\lambda,t)/c_{\alpha,\beta}(-\lambda)\big)\,dt$$

pour $\lambda\in\mathbb{C}$, $c_{\alpha,\beta}(-\lambda)$ étant la fonction de Harish-Chandra. Lorsque $\alpha=\beta=-\frac{1}{2}$ nous retrouvons la transformation de Laplace usuelle.

PLAN DE L'ARTICLE

1. Rappels et notations.
2. La transformation de Fourier–Jacobi.
3. Des transformations intégrales fractionnaires.
4. La transformation de Laplace–Jacobi.
5. La formule d'inversion.
6. Transformations de Bessel, opérateurs de Chébli.
7. Transformations radiales sur certains groupes de Lie semi-simples.
8. Eléments pour une interprétation géométrique: Noyau de Poisson hyperbolique; formule d'addition.

* Received by the editors July 17, 1981, and in revised form May 28, 1982.

† Département de Mathématiques, Université Claude Bernard, Lyon 1, 69622 Villeurbanne Cedex, France.

1. Rappels et notations. Soit $\mathbb{R}$ le corps des nombres réels, $\mathbb{C}$ le corps des complexes, $\mathbb{N}$ l'ensemble des entiers, $\mathbb{R}_+=\{t\in\mathbb{R}/t\geq 0\}$, $\mathbb{R}_+^*=\{t\in\mathbb{R}/t>0\}$ et $\mathbb{N}^*=\{n\in\mathbb{N}/n>0\}$. Tous les espaces de fonctions sont à valeurs dans $\mathbb{C}$.

Pour tout $\alpha,\beta\in\mathbb{C}$, soit la mesure $\omega_{\alpha,\beta}(t)\,dt=(\operatorname{sh}t)^{2\alpha+1}(\operatorname{ch}t)^{2\beta+1}dt$ sur $\mathbb{R}_+^*$.

Considérons l'opérateur $\Delta_{\alpha,\beta}$ formé à partir de $\omega_{\alpha,\beta}$ par:

$$(1)\qquad \Delta_{\alpha,\beta}=\frac{1}{\omega_{\alpha,\beta}(t)}\frac{d}{dt}\left(\omega_{\alpha,\beta}(t)\frac{d}{dt}\right)=\frac{d^2}{dt^2}+\left[(2\alpha+1)\coth t+(2\beta+1)\operatorname{th}t\right]\frac{d}{dt}.$$

Posons $\rho=\alpha+\beta+1$; soit pour tout $\lambda\in\mathbb{C}$ la solution $t\to\varphi_{\alpha,\beta}(\lambda,t)$ de l'opérateur

$$(2)\qquad \left(\Delta_{\alpha,\beta}+\lambda^2+\rho^2\right)f=0,$$

solution telle que $\varphi_{\alpha,\beta}(\lambda,0)=1$ et $\frac{d}{dt}\varphi_{\alpha,\beta}(\lambda,t)|_{t=0}=0$.

Pour $\alpha\neq -1,-2,\cdots$ la fonction $\varphi_{\alpha,\beta}(\lambda,\cdot)$ s'exprime à l'aide de la fonction hypergéométrique:

$$(3)\qquad \varphi_{\alpha,\beta}(\lambda,t)={}_2F_1\left(\frac{\rho+i\lambda}{2},\frac{\rho-i\lambda}{2};\alpha+1;-\operatorname{sh}^2t\right).$$

De plus, si $\lambda\neq -i,-2i,\cdots$, il existe une deuxième solution de (2) linéairement indépendante de (3), définie par la condition asymptotique suivante:

$$\Phi_{\alpha,\beta}(\lambda,t)=e^{(i\lambda-\rho)t}(1+\varepsilon(t))\text{ avec }\varepsilon(t)\to 0\text{ lorsque }t\to+\infty.$$

Cette solution s'écrit à l'aide de la fonction hypergéométrique:

$$(4)\qquad \Phi_{\alpha,\beta}(\lambda,t)=(e^t-e^{-t})^{i\lambda-\rho}\,{}_2F_1\left(\frac{\beta-\alpha+1-i\lambda}{2},\frac{\alpha+\beta+1-i\lambda}{2};1-i\lambda;\frac{-1}{\operatorname{sh}^2t}\right).$$

Les fonctions $\varphi_{\alpha,\beta}(\lambda,\cdot)$ et $\Phi_{\alpha,\beta}(\lambda,\cdot)$ sont appelées *fonctions de Jacobi* de 1-ère et de 2-ème espèce et nous avons pour tout $\alpha,\beta\in\mathbb{C}$, tout $\lambda\in\mathbb{C}$, $i\lambda\notin\mathbb{Z}$, et tout $t\in\mathbb{R}_+$:

$$(5)\qquad \frac{2\sqrt{\pi}}{\Gamma(\alpha+1)}\varphi_{\alpha,\beta}(\lambda,t)=c_{\alpha,\beta}(\lambda)\Phi_{\alpha,\beta}(\lambda,t)+c_{\alpha,\beta}(-\lambda)\Phi_{\alpha,\beta}(-\lambda,t),$$

où

$$(6)\qquad c_{\alpha,\beta}(\lambda)=2^{\rho}\frac{\Gamma\left(\frac{i\lambda}{2}\right)\Gamma\left(\frac{1+i\lambda}{2}\right)}{\Gamma\left(\frac{\alpha+\beta+1+i\lambda}{2}\right)\Gamma\left(\frac{\alpha-\beta+1+i\lambda}{2}\right)}$$

est la fonction c de Harish-Chandra.

En particulier, nous avons:

$$(7)\qquad c_{\alpha,-1/2}(\lambda)=c_{\alpha,\alpha}(2\lambda)=\frac{2^{2\alpha+1}\Gamma(i\lambda)}{\Gamma\left(\alpha+\frac{1}{2}+i\lambda\right)},$$

$$\Phi_{\alpha,-1/2}(\lambda,2t)=\Phi_{\alpha,\alpha}(2\lambda,t),\qquad \varphi_{\alpha,-1/2}(\lambda,2t)=\varphi_{\alpha,\alpha}(2\lambda,t)$$

et

$$(8)\qquad c_{-1/2,-1/2}(\lambda)=1,\qquad \omega_{-1/2,-1/2}(t)=1$$

$$\Phi_{-1/2,-1/2}(\lambda,t)=e^{i\lambda t},\qquad \varphi_{-1/2,-1/2}(\lambda,t)=\cos\lambda t.$$

2. La transformation de Fourier–Jacobi. Lorsque $\mathrm{Re}(\alpha)>-1$, pour chaque $\lambda\in\mathbb{R}_+$, et chaque $f\in L^1(\mathbb{R}_+,\omega_{\alpha,\beta}(t)\,dt)$, la transformation de Fourier–Jacobi est définie par:

$$\mathcal{F}_{\alpha,\beta}(f)(\lambda)=\hat{f}(\lambda)=\frac{2^{2\rho}\sqrt{2}}{\Gamma(\alpha+1)}\int_0^\infty f(t)\varphi_{\alpha,\beta}(\lambda,t)\omega_{\alpha,\beta}(t)\,dt. \tag{9}$$

Lorsque f est de classe $\mathcal{C}^\infty$ et à support compact sur $\mathbb{R}_+$, alors $\mathcal{F}_{\alpha,\beta}(f)(\lambda)$ se prolonge en une fonction analytique en α,β et λ. Lorsque $\mathrm{Re}(\alpha)>-n-1$ l'expression de ce prolongement analytique est donnée par [9, la formule (3.3)].

Cas particuliers. a) $\alpha=\beta=-\frac{1}{2}$; $\mathcal{F}_{-1/2,-1/2}(f)(\lambda)=\sqrt{\frac{2}{\pi}}\int_0^\infty f(t)\cos\lambda t\,dt$ est la transformation de Fourier en cosinus.

b) $\beta=-\frac{1}{2}$, $\alpha=(n-2)/2$ avec $n\in\mathbb{N}^*$; $\mathcal{F}_{(n-2)/2,-1/2}$ est la transformation de Fourier sphérique associée au groupe $SO_0(n,1)$.

c) $\beta=0$, $\alpha=n-1$ avec $n\in\mathbb{N}^*$; $\mathcal{F}_{n-1,0}$ est la transformation de Fourier sphérique associée au groupe $SU(n,1)$.

d) Pour les autres groupes de Lie-semi-simples non compacts, de centre fini et de rang réel 1, nous avons la transformation de Fourier sphérique sur $Sp(n,1)$, avec $n=2,3,4,\cdots$, en prenant $\alpha=2\,n-1$ et $\beta=1$; enfin pour $\alpha=7$ et $\beta=3$ nous avons la transformation associée au groupe exceptionnel $F_4(-20)$.

Les résultats classiques sur ces transformations de Fourier sphériques se prolongent aux transformations de Fourier–Jacobi et peuvent s'énoncer ainsi (cf. [9]):

Formule d'inversion. Lorsque $\mathrm{Re}(\alpha)>-\frac{1}{2}$, et $|\mathrm{Re}(\beta)|<\mathrm{Re}(\alpha+1), g\in L^1(\mathbb{R}_+,(c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda))^{-1}d\lambda)$,

$$\mathcal{F}_{\alpha,\beta}^{-1}(g)(t)=\frac{\sqrt{2}}{\Gamma(\alpha+1)}\int_0^\infty g(\lambda)\varphi_{\alpha,\beta}(\lambda,t)\big(c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda)\big)^{-1}d\lambda. \tag{10}$$

Formule de Bessel–Parseval. Lorsque α et $\beta\in\mathbb{R}$, $|\beta|<\alpha+1$ alors la transformation de Fourier–Jacobi se prolonge en un isomorphisme isométrique de $L^2(\mathbb{R}_+,2^{2\rho}\omega_{\alpha,\beta}(t)\,dt)$ sur $L^2(\mathbb{R}_+,|c_{\alpha,\beta}(\lambda)|^{-2}d\lambda)$ et pour tout $f,g\in L^2(\mathbb{R}_+,2^{2\rho}\omega_{\alpha,\beta}(t)\,dt)$, on *a*:

$$\int_0^\infty f(t)\overline{g(t)}2^{2\rho}\omega_{\alpha,\beta}(t)\,dt=\int_0^\infty \hat{f}(\lambda)\,\overline{\hat{g}(\lambda)}|c_{\alpha,\beta}(\lambda)|^{-2}d\lambda. \tag{11}$$

Remarque. Plus que les résultats, ce sont les méthodes de démonstrations qui nous intéressent. En voici les éléments essentiels.

Pour $\mu\in\mathbb{C}$, $\mathrm{Re}(\mu)>0$, pour $\sigma>0$ et $s\geq 0$, posons:

$$\mathcal{W}_\mu^\sigma(f)(s)=\frac{1}{\Gamma(\mu)}\int_s^\infty \frac{f(t)\,\sigma\,\mathrm{sh}\,\sigma t\,dt}{(\mathrm{ch}\,\sigma t-\mathrm{ch}\,\sigma s)^{1-\mu}} \tag{12}$$

lorsque f est de classe $\mathcal{C}^\infty$ à support compact dans $\mathbb{R}_+$.

Le prolongement analytique en μ de (12) est donné par

$$\mathcal{W}_\mu^\sigma(f)(s)=\frac{(-1)^n}{\Gamma(\mu+n)}\int_s^\infty \frac{d^n}{d(\mathrm{ch}\,\sigma t)^n}(f)(t)\frac{\sigma\,\mathrm{sh}\,\sigma t\,dt}{(\mathrm{ch}\,\sigma t-\mathrm{ch}\,\sigma s)^{1-\mu-n}} \tag{13}$$

lorsque $\mathrm{Re}(\mu)>-n$. $\mathcal{W}_\mu^1$ est en fait la transformation intégrale de Weyl (cf. [5, Chap. 13]) et nous avons:

$$\mathcal{W}_0^\sigma=\mathrm{id},\qquad \mathcal{W}_{\mu+\nu}^\sigma=\mathcal{W}_\mu^\sigma\circ\mathcal{W}_\nu^\sigma, \tag{14}$$

$$\mathcal{W}_{-1}^\sigma=-\frac{dt}{d\,\mathrm{ch}\,\sigma t}$$

pour tout $\mu,\nu\in\mathbb{C}$.

De plus, $\mathcal{W}_\mu^\sigma$ est une bijection de l'ensemble des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$, à support compact et paires, sur lui-même.

A l'aide de ces transformations intégrales, T. Koornwinder [9] établit les résultats suivants:

$$\mathcal{F}_{\alpha,\beta}(f)=2^{3\alpha+3/2}\mathcal{F}_{-1/2,-1/2}\circ\mathcal{W}_{\alpha-\beta}^1\circ\mathcal{W}_{\beta+1/2}^2(f), \tag{15}$$

formule valable pour tout $\alpha,\beta\in\mathbb{C}$ lorsque f est $\mathcal{C}^\infty$, à support compact sur $\mathbb{R}_+$. Cette formule est établie à partir des trois formules suivantes valables pour $\mathrm{Re}(\alpha)>\mathrm{Re}(\beta)>-\frac{1}{2}$ et pour f de classe $\mathcal{C}^\infty$ à support compact:

$$\text{(i)}\quad 2^{3\alpha+3/2}\mathcal{W}_{\alpha-\beta}^1\circ\mathcal{W}_{\beta+1/2}^2(f)(s)=\int_s^\infty f(t)A_{\alpha,\beta}(s,t)\,dt, \tag{16}$$

$$\text{(ii)}\quad 2^{2\rho}\omega_{\alpha,\beta}(t)\varphi_{\alpha,\beta}(\lambda,t)=\frac{\Gamma(\alpha+1)}{\sqrt{\pi}}\int_0^t\cos\lambda s A_{\alpha,\beta}(s,t)\,ds,$$

$$\text{(iii)}\quad A_{\alpha,\beta}(s,t)=\frac{2^{3(\alpha+1/2)}2\,\mathrm{sh}\,2t}{\Gamma(\alpha-\beta)\Gamma(\beta+\frac{1}{2})}\int_s^t\frac{(\mathrm{ch}\,u-\mathrm{ch}\,s)^{\alpha-\beta-1}}{(\mathrm{ch}\,2t-\mathrm{ch}\,2u)^{1/2-\beta}}\mathrm{sh}\,u\,du$$

$$=\frac{2^{\alpha+2\beta+3/2}}{\Gamma(\alpha+\frac{1}{2})}2\,\mathrm{sh}\,2t(\mathrm{ch}\,t)^{\beta-\alpha}(\mathrm{ch}\,2t-\mathrm{ch}\,2s)^{\alpha-1/2}$$

$$\times{}_2F_1\left(\alpha+\beta,\alpha-\beta;\alpha+\frac{1}{2};\frac{\mathrm{ch}\,t-\mathrm{ch}\,s}{2\,\mathrm{ch}\,t}\right).$$

A partir de la formule (13) est démontré le théorème de Paley–Wiener; cf. [9, théorèmes (3.4) et (4.2)].

Soit $\mathcal{C}_0^\infty(\mathbb{R})$ l'ensemble des fonctions paires à support compact sur $\mathbb{R}$ et de classe $\mathcal{C}^\infty$. Soit $\mathcal{H}$ l'ensemble des fonctions analytiques sur $\mathbb{C}$, paires et rapidement décroissantes, de type exponentiel.

THÉORÈME. (i) *Pour tout $\alpha,\beta\in\mathbb{C}$, $\mathcal{F}_{\alpha,\beta}$ est une bijection de $\mathcal{C}_0^\infty(\mathbb{R})$ sur $\mathcal{H}$.*

(ii) *L'application réciproque est donée par*

$$\mathcal{F}_{\alpha,\beta}^{-1}(g)(t)=\frac{\sqrt{2}}{\Gamma(\alpha+1)}\int_0^\infty\frac{g(\lambda)\varphi_{\alpha,\beta}(\lambda,t)}{c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda)}d\lambda$$

lorsque $|\mathrm{Re}(\beta)|<\mathrm{Re}(\alpha+1)$, *et* $\mathrm{Re}(\alpha)>-\frac{1}{2}$.

3. Un outil pratique: Les "transformations intégrales fractionnaires". En considérant les résultats exposés par Viano [14], à propos d'une transformation de Laplace liée au groupe $SL(2,\mathbb{R})$ (i.e. $\alpha=0$ et $\beta=-\frac{1}{2}$) et en utilisant les méthodes brièvement rappelées ci-dessus, nous pouvons définir une transformation intégrale attachée à la fonction de Jacobi de 2-ème espèce.

Soit $\lambda\in\mathbb{C}$ et f une fonction de classe $\mathcal{C}^\infty$ et à support compact dans $\mathbb{R}_+$; les calculs montrent rapidement qu'au lieu de définir cette transformation par la formule

$2^{2\rho}\sqrt{2}/\Gamma(\alpha+1)\int_0^\infty f(t)\Phi_{\alpha,\beta}(\lambda,t)\omega_{\alpha,\beta}(t)\,dt$, (i.e. en recopiant la définition de la transformation de Fourier–Jacobi), il vaut mieux poser:

$$\mathcal{L}_{\alpha,\beta}(f)(\lambda)=\tilde{f}(\lambda)=\int_0^\infty f(t)\,\frac{\Phi_{\alpha,\beta}(\lambda,t)}{c_{\alpha,\beta}(-\lambda)}\,dt. \tag{17}$$

Pour faciliter l'exposition des résultats, il nous faut d'abord introduire une transformation intégrale de type transformation intégrale fractionnaire de Riemann–Liouville; cf. [5, Chap. 13]; en plus de la transformation du type transformation intégrale fractionnaire de Weyl, cf. formules (12), (13) et (14).

Soit $\delta>0$ et $\mathcal{C}_\delta$ l'ensemble des fonctions continues, à support inclus dans $[\delta,+\infty[$; nous noterons $\mathcal{C}_\delta^\infty$ l'ensemble des fonctions qui sont de plus de classe $\mathcal{C}^\infty$ sur $\mathbb{R}_+$.

Soit $\sigma>0$, $t\geq 0$ et $f\in\mathcal{C}_\delta$; posons pour $\mu\in\mathbb{C}$, $\mathrm{Re}(\mu)>0$:

$$\mathcal{R}_\mu^\sigma(f)(t)=\frac{1}{\Gamma(\mu)}\int_0^t f(s)\,\frac{d\,\mathrm{ch}\,\sigma s}{(\mathrm{ch}\,\sigma t-\mathrm{ch}\,\sigma s)^{1-\mu}}. \tag{18}$$

L'application $\mu\to\mathcal{R}_\mu^\sigma(f)(t)$ est holomorphe sur $\mathrm{Re}(\mu)>0$ et si de plus $f\in\mathcal{C}_\delta^\infty$ elle admet un prolongement analytique à tout le plan complexe défini par:

(19)

$$\mathcal{R}_\mu^\sigma(f)(t)=\frac{1}{\Gamma(\mu+n)}\int_0^t\frac{d^nf}{d(\mathrm{ch}\,\sigma s)^n}(s)\,\frac{d\,\mathrm{ch}\,\sigma s}{(\mathrm{ch}\,\sigma t-\mathrm{ch}\,\sigma s)^{1-\mu-n}}\quad\text{dès que } \mathrm{Re}(\mu+n)>0.$$

Lorsque f est une fonction continue sur $\mathbb{R}_+$, la formule (18) garde un sens; mais lorsque f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}_+$, l'expression (19) donne un prolongement analytique de (18) sur le demi-plan $\mathrm{Re}(\mu+n)>0$ si pour tout k entier variant de 1 à n, $\lim_{t\to 0}(d^k/d(\mathrm{ch}\,\sigma t)^k)f(t)=0$.

De plus, on a les formules suivantes:

$$\mathcal{R}_0^\sigma=\mathrm{id},\qquad \mathcal{R}_{\mu+\nu}^\sigma=\mathcal{R}_\mu^\sigma\circ\mathcal{R}_\nu^\sigma\quad\text{pour tout }\mu,\nu\in\mathbb{C}, \tag{20}$$

$$\mathcal{R}_\mu^\sigma\left(\frac{d}{d\,\mathrm{ch}\,\sigma t}f\right)(s)=\frac{d}{d\,\mathrm{ch}\,\sigma s}\mathcal{R}_\mu^\sigma(f)(s)=\mathcal{R}_{\mu-1}^\sigma(f)(s)\quad\text{pour tout }\mu\in\mathbb{C},\ s\geq 0$$

et tout $f\in\mathcal{C}_\delta^\infty$.

Lorsque f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}_+$, $\mathcal{R}_{-1}^\sigma(f)(s)=(d/d\,\mathrm{ch}\,\sigma s)f(s)+f(0)T$, où T est une distribution de support $\{0\}$. Ceci provient du fait que

$$\mathcal{R}_{-1}^\sigma(f)(t)=\lim_{\mu\to -1}\mathcal{R}_\mu^\sigma(f)(t)=f(0)\frac{(\mathrm{ch}\,\sigma t-1)^\mu}{\Gamma(\mu+1)}+\frac{df}{d\,\mathrm{ch}\,\sigma s}(0)\frac{(\mathrm{ch}\,\sigma t-1)^{\mu+1}}{\Gamma(\mu+2)}$$
$$+\frac{1}{\Gamma(\mu+1)}\int_0^t\frac{d^2f}{d(\mathrm{ch}\,\sigma s)^2}(s)(\mathrm{ch}\,\sigma t-\mathrm{ch}\,\sigma s)^{\mu+1}d\,\mathrm{ch}\,\sigma s.$$

Précisons enfin que pour tout $\mu\in\mathbb{C}$ et tout $\sigma>0$, $\mathcal{R}_\mu^\sigma$ est une bijection de $\mathcal{C}_\delta^\infty$ sur $\mathcal{C}_\delta^\infty$.

Ecrivons en utilisant les transformations $\mathcal{W}_\mu^\sigma$ et $\mathcal{R}_\mu^\sigma$ les relations fondamentales liant les fonctions de Jacobi entre elles:

a) La formule [9, (2.14)] valable pour $t\in\mathbb{R}_+$, $\beta\in\mathbb{C}$, $\lambda\in\mathbb{C}$, $\mathrm{Re}(\alpha)>-1$ et $\mathrm{Re}(\mu)>0$ s'écrit

$$\frac{\omega_{\alpha+\mu,\beta+\mu}(t)\varphi_{\alpha+\mu,\beta+\mu}(\lambda,t)}{\Gamma(\alpha+\mu+1)\,\mathrm{sh}\,2t}=2^{-\mu}\mathcal{R}_\mu^2\left(\frac{\omega_{\alpha,\beta}(\cdot)\varphi_{\alpha,\beta}(\lambda,\cdot)}{\Gamma(\alpha+1)\,\mathrm{sh}\,2\cdot}\right)(t) \tag{21}$$

et s'étend, par exemple, à $\alpha\in\mathbb{C}$, $\beta\in\mathbb{C}$, $\lambda\in\mathbb{C}$, $\mu\in\mathbb{C}$, $t\in\mathbb{R}_+$ avec $\mathrm{Re}(\mu)>-\mathrm{Re}(\alpha)$ et $-(\alpha+1)\notin\mathbb{N}$.

b) La formule [9, (2.15)] valable pour $t\in\mathbb{R}_+^*$, $\alpha\in\mathbb{C}$, $\beta\in\mathbb{C}$, $\lambda\in\mathbb{C}$, $\mu\in\mathbb{C}$ avec $\mathrm{Re}(\mu)>0$ et $\mathrm{Im}(\lambda)>-\mathrm{Re}(\alpha+\beta+1)+2\,\mathrm{Re}(\mu)$ s'écrit:

$$\frac{\Phi_{\alpha-\mu,\beta-\mu}(\lambda,t)}{c_{\alpha-\mu,\beta-\mu}(-\lambda)}=2^{3\mu}\mathcal{W}_\mu^2\left(\frac{\Phi_{\alpha,\beta}(\lambda,\cdot)}{c_{\alpha,\beta}(-\lambda)}\right)(t) \tag{22}$$

et est prolongeable à tout $\mu\in\mathbb{C}$, dès que $\mathrm{Im}(\lambda)>-\mathrm{Re}(\alpha+\beta+1)+2\,\mathrm{Re}(\mu)$. Cette formule (22) permet d'obtenir, en tenant compte de (5), "une formule duale" de (21):

(23)

$$\varphi_{\alpha,\beta}(\lambda,t)=2^{3\mu}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\mu+1)}\frac{c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda)}{c_{\alpha+\mu,\beta+\mu}(\lambda)c_{\alpha+\mu,\beta+\mu}(-\lambda)}\mathcal{W}_\mu^2\big(\varphi_{\alpha+\mu,\beta+\mu}(\lambda,\cdot)\big)(t),$$

lorsque $|\mathrm{Im}(\lambda)|<\mathrm{Re}(\alpha+\beta+1)$.

Cas particulier. Lorsque $\beta=-\frac{1}{2}$, les formules ci-dessus deviennent, en utilisant (7) et (5),

$$\varphi_{\alpha,-1/2}(\lambda,t)=\frac{2^{\alpha+1/2}\Gamma(\alpha+1)}{\sqrt{\pi}\,\omega_{\alpha,-1/2}(t)}\,\mathrm{sh}\,t\,\mathcal{R}^1_{\alpha+1/2}\left(\frac{\cos\lambda\,\cdot}{\mathrm{sh}\,\cdot}\right)(t) \tag{24}$$

et, lorsque $\mathrm{Im}(\lambda)>-\mathrm{Re}(\alpha+\frac{1}{2})$,

$$\varphi_{\alpha,-1/2}(\lambda,t)=2^{-3(\alpha+1/2)}c_{\alpha,-1/2}(-\lambda)\mathcal{W}^1_{-(\alpha+1/2)}(e^{i\lambda\cdot})(t). \tag{25}$$

Remarquons que les formules (21) et (24) permettent de réécrire (16)(ii) à l'aide des transformations intégrales fractionnaires:

(26)

$$2^{2\rho}\omega_{\alpha,\beta}(t)\varphi_{\alpha,\beta}(\lambda,t)=2^{3(\alpha+1/2)}\frac{\Gamma(\alpha+1)}{\sqrt{\pi}}\,\mathrm{sh}\,2t\left[\mathcal{R}^2_{\beta+1/2}\frac{1}{2\,\mathrm{ch}\,s}\left(\mathcal{R}^1_{\alpha-\beta}\left(\frac{\cos\lambda\,\cdot}{\mathrm{sh}\,\cdot}\right)\right)(s)\right](t).$$

De même à l'aide de la formule (23) et de la formule suivante obtenue de manière similaire lorsque $\beta=-\frac{1}{2}$:

$$\varphi_{\alpha,-1/2}(\lambda,t)=\frac{2^{-3(\alpha+1/2)}\Gamma(\alpha+1)}{\sqrt{\pi}}c_{\alpha,-1/2}(\lambda)c_{\alpha,-1/2}(-\lambda)\mathcal{W}^1_{-(\alpha+1/2)}(\cos\lambda\,\cdot)(t)$$

on obtient une "formule duale" de la formule (26), valable lorsque $|\mathrm{Im}(\lambda)|<\mathrm{Re}(\alpha+\beta+1)$:

$$\frac{\varphi_{\alpha,\beta}(\lambda,t)}{c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda)}=\frac{2^{-3(\alpha+1/2)}\Gamma(\alpha+1)}{\sqrt{\pi}}\mathcal{W}^2_{-(\beta+1/2)}\circ\mathcal{W}^1_{\beta-\alpha}(\cos\lambda\,\cdot)(t).$$

(Pour $t=0$ cette formule nous donne une nouvelle évaluation de la mesure de Plancherel.) Enfin, à l'aide des formules (22) et (25) on retrouve dans une autre forme la formule [9, (2.17)]:

$$\frac{\Phi_{\alpha,\beta}(\lambda,t)}{c_{\alpha,\beta}(-\lambda)}=2^{-3(\alpha+1/2)}\mathcal{W}^2_{-\beta-1/2}\circ\mathcal{W}^1_{\beta-\alpha}(e^{i\lambda\cdot})(t) \tag{27}$$

$$\text{lorsque}\ \mathrm{Im}(\lambda)>-\mathrm{Re}(\alpha+\beta+1),\quad \mathrm{Im}(\lambda)>\mathrm{Re}(\beta-\alpha).$$

C'est cette dernière formule qui est le point de départ de la transformation de Laplace–Jacobi, objet de cet article.

4. La transformation de Laplace–Jacobi.

PROPOSITION 4.1. *Soit f une fonction de classe $\mathcal{C}^\infty$ à support compact dans $\mathbb{R}_+^*$, soit $\alpha,\beta,\lambda\in\mathbb{C}$; alors, pour* $\mathrm{Im}(\lambda)>\mathrm{Max}(-\mathrm{Re}(\alpha+\beta+1),\mathrm{Re}(\beta-\alpha))$, *on a*

$$(28)\quad \mathcal{L}_{\alpha,\beta}(f)(\lambda)=2^{-3(\alpha+1/2)}\mathcal{L}_{-1/2,-1/2}\left[\mathrm{sh}\,\mathcal{R}^1_{\beta-\alpha}\circ\mathrm{ch}\,\mathcal{R}^2_{-\beta-1/2}\left(\frac{f}{\mathrm{sh}\,\mathrm{ch}}\right)\right](\lambda).$$

La considération des pôles de $c_{\alpha,\beta}(-\lambda)^{-1}$ permet de dire que $(\alpha,\beta,\lambda)\to\Phi_{\alpha,\beta}(\lambda,t)/c_{\alpha,\beta}(-\lambda)$ est holomorphe pour tout $t>0$ dans la région $\mathrm{Im}(\lambda)>\mathrm{Max}(-\mathrm{Re}(\alpha+\beta+1);\ \mathrm{Re}(\beta-\alpha-1))$ ainsi en utilisant la formule (27) après permutations d'intégrales, on a le résultat.

De plus, la formule (28) a un sens pour $\mathrm{Im}(\lambda)>-\mathrm{Re}(\alpha+\beta+1)$ et la formule (17) pour $\frac{1}{2}(\alpha+\beta+1-i\lambda)\notin-\mathbb{N}$ et $\frac{1}{2}(\alpha-\beta+1-i\lambda)\notin-\mathbb{N}$. Les formules (17) et (28) constituent des prolongements analytiques l'une de l'autre.

Remarque 4.2. i) $\mathcal{L}_{-1/2,-1/2}$ est la transformation de Laplace usuelle pour la variable $-i\lambda$. Ainsi, de même que la transformation de Fourier–Jacobi $\mathcal{F}_{\alpha,\beta}$ s'interprète comme une généralisation de la transformation de Fourier en cosinus, la transformation $\mathcal{L}_{\alpha,\beta}$ s'interprète comme généralisation de la transformation de Laplace.

ii) En utilisant les formules (5), (9) et (17) on a, pour f de classe $\mathcal{C}^\infty$ à support compact dans $\mathbb{R}_+^*$, pour $\alpha,\beta\in\mathbb{C}$, $\mathrm{Re}(\alpha+\beta+1)>0$ et pour $\lambda\in\mathbb{C}$, $|\mathrm{Im}(\lambda)|<\mathrm{Re}(\alpha+\beta+1)$,

$$(29)\quad \frac{1}{c_{\alpha,\beta}(\lambda)c_{\alpha,\beta}(-\lambda)}\mathcal{F}_{\alpha,\beta}(f)(\lambda)=\frac{2^{2\rho}}{\sqrt{2\pi}}\left\{\mathcal{L}_{\alpha,\beta}\big(f(\cdot)\omega_{\alpha,\beta}(\cdot)\big)(\lambda)+\mathcal{L}_{\alpha,\beta}\big(f(\cdot)\omega_{\alpha,\beta}(\cdot)\big)(-\lambda)\right\}.$$

De plus, pour $\mathrm{Re}(\alpha)>-\frac{1}{2}$ et $|\mathrm{Re}(\beta)|<\mathrm{Re}(\alpha+1)$, on obtient une première formule d'inversion:

$$f(t)=\frac{2^{2\rho}}{\Gamma(\alpha+1)\sqrt{\pi}}\int_{-\infty}^{+\infty}\mathcal{L}_{\alpha,\beta}\big(f(\cdot)\omega_{\alpha,\beta}(\cdot)\big)(\lambda)\varphi_{\alpha,\beta}(\lambda,t)\,d\lambda.$$

LEMMA 4.3. (cf. [9, Lemmes (2.1) et (2.2)]).

(i) *Pour tout $\alpha,\beta\in\mathbb{C}$, $\alpha\notin-\mathbb{N}^*$, pour tout $\delta>0$, il existe une constante $K_1>0$ telle que pour tout $t>\delta$ et tout $\lambda\in\mathbb{C}$, $\mathrm{Im}(\lambda)\geq0$, on a*:

$$\left|\Phi_{\alpha,\beta}(\lambda,t)\right|\leq K_1 e^{-(\mathrm{Im}(\lambda)+\mathrm{Re}(\rho))t}.$$

(ii) *Pour tout $\alpha,\beta\in\mathbb{C}$, pour tout $r>0$, il existe une constante $K_2>0$ telle que, pour tout $\lambda\in\mathbb{C}$, $\mathrm{Im}(\lambda)\geq0$, $\mathrm{Im}(\lambda)\geq-\mathrm{Re}(\rho)+r$ et $\mathrm{Im}(\lambda)\geq-\mathrm{Re}(\alpha-\beta+1)+r$, on a*:

$$\left|\frac{1}{c_{\alpha,\beta}(-\lambda)}\right|\leq K_2\big(1+|\lambda|\big)^{\mathrm{Re}(\alpha)+1/2}.$$

Pour $a\in\mathbb{R}$, soit $\mathcal{C}_{\delta,a}=\{f\in\mathcal{C}_\delta\,|\,\exists K>0,\ |f(t)|\leq Ke^{at}$ pour tout $t>0\}$, et soit $\mathcal{C}^\infty_{\delta,a}=\{f\in\mathcal{C}_\delta\,|\,\forall k\in\mathbb{N},\ (\frac{d}{dt})^n f\in\mathcal{C}_{\delta,a}\}$.

LEMME 4.4. *Soit $f\in\mathcal{C}^\infty_{\delta,a}$ et soit $\mu\in\mathbb{C}$, alors $\mathcal{R}^\sigma_\mu(f)\in\mathcal{C}^\infty_{\delta,a+\mathrm{Re}(\mu)}$ pour $a>0$; c'est-à-dire $\mathcal{R}^\sigma_\mu$ est une bijection de $\mathcal{C}^\infty_{\delta,a}$ sur $\mathcal{C}^\infty_{\delta,a+\sigma\,\mathrm{Re}(\mu)}$ lorsque $a>0$.*

En utilisant le fait que pour $\mu,\nu\in\mathbb{C}$, $\mathrm{Re}(\mu)>0$, $\mathrm{Re}(\nu)>0$ et $t>0$: $\mathcal{R}^{\sigma}_{\mu}((\mathrm{ch}\,\sigma s)^{\nu-1})(t)=(\Gamma(\nu)/\Gamma(\mu+\nu))[(\mathrm{ch}\,\sigma t)^{\mu+\nu-1}-1]$, cf. [5, 13.1.(7)], on obtient, pour $a>-\sigma$, $\mathrm{Re}(\mu)>0$ et $f\in\mathcal{C}_{\delta,a}$: $\mathcal{R}^{\sigma}_{\mu}(f)\in\mathcal{C}_{\delta,a+\sigma\,\mathrm{Re}(\mu)}$.

En remarquant ensuite que pour $r\in\mathbb{R}$ et $f\in\mathcal{C}_{\delta,a}$ on a $e^{rt}f\in\mathcal{C}_{\delta,a+r}$, et du fait que pour $f\in\mathcal{C}^{\infty}_{\delta}$ on a

$$\frac{d}{dt}\mathcal{R}^{\sigma}_{\mu}(f)(t)=\mathrm{sh}\,\sigma t\,\mathcal{R}^{\sigma}_{\mu}\left(\frac{1}{\mathrm{sh}\,\sigma s}\frac{d}{ds}f(s)\right)(t),$$

alors pour $f\in\mathcal{C}^{\infty}_{\delta,a}$ on a $\mathcal{R}^{\sigma}_{\mu}(f)\in\mathcal{C}^{\infty}_{\delta,a+\sigma\,\mathrm{Re}(\mu)}$ lorsque $a>0$ et $\mathrm{Re}(\mu)>0$. Enfin pour $\mu\in\mathbb{C}$ on utilise le fait que si $f\in\mathcal{C}^{\infty}_{\delta,a}$ alors $\mathcal{R}^{\sigma}_{-1}(f)=(d/d\,\mathrm{ch}\,\sigma t)f\in\mathcal{C}^{\infty}_{\delta,a-\sigma}$.

Remarque. Pour $a\leq 0$, on obtient pour $f\in\mathcal{C}^{\infty}_{\delta,a}$, $\mathcal{R}^{\sigma}_{\mu}(f)\in\mathcal{C}^{\infty}_{\delta,\mathrm{Max}(0,a+\sigma\,\mathrm{Re}(\mu))}$.

PROPOSITION 4.5. *Soit* $\alpha,\beta\in\mathbb{C}$, $\lambda\in\mathbb{C}$ *et* $a\in\mathbb{R}$.

(i) *Soit* $f\in\mathcal{C}_{\delta,a}$; *alors* $\lambda\mapsto\mathcal{L}_{\alpha,\beta}(f)(\lambda)$ *est holomorphe dans le demi-plan* $\mathrm{Im}(\lambda)>b=\mathrm{Max}(a-Re(\rho),0,-\mathrm{Re}(\rho),-\mathrm{Re}(\alpha-\beta+1))$ *et pour tout* $r>0$ *il existe* $K>0$ *tel que*

$$|\mathcal{L}_{\alpha,\beta}(f)(\lambda)|\leq K(1+|\lambda|)^{\mathrm{Re}(\alpha+1/2)}e^{-\delta\,\mathrm{Im}(\lambda)}\quad \textit{pour tout } \lambda,\ \mathrm{Im}(\lambda)\geq b+r.$$

(ii) *Soit* $a>0$ *et* $f\in\mathcal{C}^{\infty}_{\delta,a}$; *alors* $t\to\mathrm{sh}\,t[\mathcal{R}^{1}_{\beta-\alpha}\circ\mathrm{ch}\,\mathcal{R}^{2}_{-\beta-1/2}(f/\mathrm{sh}\,\mathrm{ch})](t)$ *appartient à* $\mathcal{C}^{\infty}_{\delta,a-\mathrm{Re}(\rho)}$; *en conséquence, pour tout* $k\in N^{*}$, *il existe* $K>0$, *tel que*

$$|\mathcal{L}_{\alpha,\beta}(f)(\lambda)|\leq\frac{K}{(1+|\lambda|)^{k}}e^{-\delta\,\mathrm{Im}(\lambda)}\quad \textit{dès que } \mathrm{Im}(\lambda)>a-\mathrm{Re}(\rho).$$

(i) est une conséquence du lemme 4.3, appliqué à la formule (17); (ii) est une conséquence du lemme 4.4, de la proposition 4.1. (formule (28)) et des propriétés élémentaires de la transformation de Laplace $\mathcal{L}_{-1/2,-1/2}$.

Remarques 4.6. (i) Soit D_2 l'opérateur défini sur $\mathcal{C}^{\infty}_{\delta,a}$ par

$$D_2(f)(t)=\frac{d}{dt}\left(\frac{f(t)}{2\,\mathrm{sh}\,2t}\right)=\mathrm{sh}\,2t\,\frac{d}{d\,\mathrm{ch}(2t)}\left(\frac{f(t)}{\mathrm{sh}\,2t}\right),$$

on a

$$D_2(f)(t)=\frac{d}{d\,\mathrm{ch}\,2t}f(t)-\frac{1}{\mathrm{sh}\,2t}\coth 2t f(t)$$

et pour tout $k\in\mathbb{N}$, on obtient $\mathcal{L}_{\alpha,\beta}(D_2^{k}(f))=2^{3k}\mathcal{L}_{\alpha+k,\beta+k}(f)$.

(ii) Considérons le cas particulier $\beta=-\frac{1}{2}$; un certain nombre de formules se simplifient, notamment

$$\mathcal{L}_{\alpha,-1/2}(f)(\lambda)=2^{-3(\alpha+1/2)}\mathcal{L}_{-1/2,-1/2}\left(\mathrm{sh}\,\mathcal{R}^{1}_{-(1/2)-\alpha}\left(\frac{f}{\mathrm{sh}}\right)\right)(\lambda).$$

Ainsi, on peut reformuler sans difficulté la proposition 4.5(ii), en considérant l'espace de fonctions $\mathcal{C}^{\infty}_{\delta,a}$. Par ailleurs, si l'on pose, pour $f\in\mathcal{C}^{\infty}_{\delta,a}$, $D_1(f)=\frac{d}{dt}(f(t)/\mathrm{sh}(t))$, on obtient, pour tout $k\in\mathbb{N}$, $\mathcal{L}_{\alpha,-1/2}\circ D_1^{k}=2^{3k}\mathcal{L}_{\alpha+k,-1/2}$.

(iii) Enfin, en considérant les formules (7) et (17), le cas particulier $\alpha=\beta$ se ramène au cas $\beta=-\frac{1}{2}$ par la formule

$$\mathcal{L}_{\alpha,\alpha}(f(t))(2\lambda)=\frac{1}{2}\mathcal{L}_{\alpha,-1/2}\left(f\left(\frac{t}{2}\right)\right)(\lambda),\quad \text{lorsque } f\in\mathcal{C}_{\delta,a}.$$

5. La formule d'inversion. Pour établir la transformation inverse de la transformation de Laplace–Jacobi, procédons formellement en inversant la formule (28).

Soit $\alpha,\beta\in\mathbb{C}$ et soit $\lambda\to g(\lambda)$ une fonction holomorphe sur un demi-plan $\mathrm{Im}(\lambda)>b$, et telle que $\frac{1}{\pi}\int_{ia-\infty}^{ia+\infty}g(\lambda)e^{-i\lambda u}d\lambda$ existe et ne dépende pas de $a>b$ pour tout $u\in\mathbb{R}_{+}$.

Nous avons alors:

$$\mathcal{L}^{-1}_{\alpha,\beta}(g)(t)=2^{3(\alpha+1/2)}\operatorname{sh}t\operatorname{ch}t\mathcal{R}^2_{\beta+1/2}\left[\frac{1}{\operatorname{ch}}\mathcal{R}^1_{\alpha-\beta}\left(\frac{1}{\operatorname{sh}}\mathcal{L}^{-1}_{-1/2,-1/2}(g)\right)\right](t).$$

C'est-à-dire si, par exemple, $\operatorname{Re}(\beta)>-\frac{1}{2}$ et $\operatorname{Re}(\alpha)>\operatorname{Re}(\beta)$,

$$\mathcal{L}^{-1}_{\alpha,\beta}(g)(t)=2^{3(\alpha+1/2)}\operatorname{sh}t\operatorname{ch}t\frac{1}{\Gamma(\beta+\frac{1}{2})}\int_0^t\frac{2\operatorname{sh}2s\,ds}{\operatorname{ch}s(\operatorname{ch}2t-\operatorname{ch}2s)^{1-\beta-1/2}}$$
$$\times\frac{1}{\Gamma(\alpha-\beta)}\int_0^s\frac{\operatorname{sh}u\,du}{\operatorname{sh}u(\operatorname{ch}s-\operatorname{ch}u)^{1-\alpha+\beta}}\times\frac{1}{\pi}\int_{ia-\infty}^{ia+\infty}g(\lambda)e^{-i\lambda u}d\lambda$$

et en intervertissant (toujours formellement) les intégrales:

$$\mathcal{L}^{-1}_{\alpha,\beta}(g)(t)=2^{3(\alpha+1/2)}\operatorname{sh}t\operatorname{ch}t\frac{1}{\pi}\int_{ia-\infty}^{ia+\infty}g(\lambda)\times\frac{1}{\Gamma(\beta+\frac{1}{2})}\int_0^t\frac{2\operatorname{sh}2s\,ds}{\operatorname{ch}s(\operatorname{ch}2t-\operatorname{ch}2s)^{1-\beta-1/2}}$$
$$\times\frac{1}{\Gamma(\alpha-\beta)}\int_0^s\frac{e^{-i\lambda u}\operatorname{sh}u\,du}{\operatorname{sh}u(\operatorname{ch}s-\operatorname{ch}u)^{1-\alpha+\beta}}.$$

Posons alors, par définition, pour $\operatorname{Re}(\alpha)>\operatorname{Re}(\beta)>-\frac{1}{2}$

$$\psi_{\alpha,\beta}(\lambda,t)=2^{3(\alpha+1/2)}\operatorname{sh}t\operatorname{ch}t\left[\mathcal{R}^2_{\beta+1/2}\frac{1}{\operatorname{ch}}\mathcal{R}^1_{\alpha-\beta}\left(\frac{e^{-i\lambda\cdot}}{\operatorname{sh}\cdot}\right)\right](t) \tag{30}$$

et

$$\check{g}(t)=\frac{1}{\pi}\int_{ia-\infty}^{ia+\infty}g(\lambda)\psi_{\alpha,\beta}(\lambda,t)\,d\lambda. \tag{31}$$

Etudions alors les conditions d'existence de ces formules (30) et (31).

PROPOSITION 5.1. *Soit* $\alpha,\beta,\lambda\in\mathbb{C}$ *et* $t\in\mathbb{R}_+$. *Lorsque* $\operatorname{Re}(\alpha)>\operatorname{Re}(\beta)>-\frac{1}{2}$, $\psi_{\alpha,\beta}(\lambda,t)=\int_0^t e^{-i\lambda s}A_{\alpha,\beta}(s,t)\,ds$, *où* $A_{\alpha,\beta}$ *est défini par la formule* (16) (iii).

De plus, il existe une constante $K\geq 0$ *telle que pour tout* λ,

$$\operatorname{Im}(\lambda)>0,\ |\psi_{\alpha,\beta}(\lambda,t)|\leq K(1+t)e^{(\operatorname{Im}(\lambda)+\operatorname{Re}(\rho))t}.$$

Cette proposition est une conséquence de la définition de $A_{\alpha,\beta}$; la majoration repose sur son expression à l'aide de la fonction hypergéométrique.

Remarques 5.2. (i) Pour $k\in\mathbb{N}$, pour $\operatorname{Re}(\alpha-k)>\operatorname{Re}(\beta-k)>-\frac{1}{2}$ on a

$$\psi_{\alpha-k,\beta-k}(\lambda,t)=2^{-3k}D_2^k\psi_{\alpha,\beta}(\lambda,t).$$

(ii) Lorsque $\alpha=\beta$ avec $\operatorname{Re}(\beta)>-\frac{1}{2}$, ou lorsque $\operatorname{Re}(\alpha)>\beta=-\frac{1}{2}$, $\psi_{\alpha,\beta}$ a un sens, par exemple

$$\psi_{\alpha,-1/2}(\lambda,t)=2^{3(\alpha+1/2)}\operatorname{sh}t\mathcal{R}^1_{\alpha+1/2}\left(\frac{e^{-i\lambda s}}{\operatorname{sh}s}\right)(t).$$

(iii) En utilisant la formule (16) (ii), on obtient la relation suivante lorsque $\operatorname{Re}(\alpha)>\operatorname{Re}(\beta)>-\frac{1}{2}$:

$$\psi_{\alpha,\beta}(\lambda,t)+\psi_{\alpha,\beta}(-\lambda,t)=\frac{2\sqrt{\pi}}{\Gamma(\alpha+1)}2^{2\rho}\omega_{\alpha,\beta}(t)\varphi_{\alpha,\beta}(\lambda,t). \tag{32}$$

(iv) En appliquant $\mathcal{R}^2_\mu$ à la formule (30), on obtient pour tout $\mu\in\mathbb{C}$, $\mathrm{Re}(\mu)>0$,

$$\frac{\psi_{\alpha+\mu,\beta+\mu}(\lambda,t)}{\mathrm{sh}\,2t}=2^{3\mu}\mathcal{R}^2_\mu\left(\frac{\psi_{\alpha,\beta}(\lambda,s)}{\mathrm{sh}\,2s}\right)(t); \tag{33}$$

de même, en appliquant $\mathcal{R}^1_\mu$ à (ii) ci-dessus, on obtient pour $\mathrm{Re}(\mu)>0$,

$$\frac{\psi_{\alpha+\mu,-1/2}(\lambda,t)}{\mathrm{sh}\,t}=2^{3\mu}\mathcal{R}^1_\mu\left(\frac{\psi_{\alpha,-1/2}(\lambda,s)}{\mathrm{sh}\,s}\right)(t). \tag{34}$$

Notations 5.3. Soit $\delta>0$ et $a\in\mathbb{R}$, soit $\mathcal{H}_{\delta,a}$ l'espace des fonctions holomorphes, $\lambda\to g(\lambda)$, sur le demi-plan $\mathrm{Im}(\lambda)>a$ et telles que pour tout $n\in\mathbb{N}$, il existe une constante $K_n>0$ pour laquelle $|g(\lambda)|\le(K_n/(1+|\lambda|)^n)e^{-\delta\mathrm{Im}(\lambda)}$.

THÉORÈME 5.4. *Soit $\alpha,\beta\in\mathbb{C}$, $a\in\mathbb{R}$ et $\delta>0$.*

(i) *Soit $a>0$, $\mathcal{L}_{\alpha,\beta}$ est une bijection de $\bigcap_{b>a}\mathcal{C}^\infty_{\delta,b}$ sur $\bigcap_{b>a}\mathcal{H}_{\delta,b-\mathrm{Re}(\rho)}$.*

(ii) *De plus, si $\mathrm{Re}(\alpha)>\mathrm{Re}(\beta)>-\frac12$, pour $a>-\mathrm{Re}(\rho)$, la bijection réciproque est donnée par $g\to\check g$ où $\check g(t)=\frac1\pi\int_{ib-\infty}^{ib+\infty}g(\lambda)\psi_{\alpha,\beta}(\lambda,t)\,d\lambda$ est indépendant de $b>a$, et pour $g\in\mathcal{H}_{\delta,a}$, $\check g\in\mathcal{C}^\infty_{\delta,a+\mathrm{Re}(\rho)+\varepsilon}$ pour tout $\varepsilon>0$.*

(i) est une conséquence de la proposition 4.5(ii) et des propriétés de la transformation de Laplace $\mathcal{L}_{-1/2,-1/2}$, en utilisant la formule (28).

(ii) Par la formule de Cauchy et la majoration de la proposition 5.1, on obtient l'indépendance de $\check g(t)$ vis à vis de $b>a$. D'autre part, comme $|\check g(t)|\le K(1+t)e^{b(t-\delta)}e^{\mathrm{Re}(\rho)t}$, lorsque $b\to+\infty$, on obtient $\check g(t)=0$ si $0\le t<\delta$ et, si b tend vers a, on obtient $|\check g(t)|\le K_1e^{(\mathrm{Re}(\rho)+a+\varepsilon)t}$ pour tout $\varepsilon>0$.

Comme $g\in\mathcal{H}_{\delta,a}$, on a également:

$$\check g(t)=2^{3(\alpha+1/2)}\,\mathrm{sh}\,t\,\mathrm{ch}\,t\,\mathcal{R}^2_{\beta+1/2}\circ\frac{1}{\mathrm{ch}}\mathcal{R}^1_{\alpha-\beta}\circ\frac{1}{\mathrm{sh}}\mathcal{L}^{-1}_{-1/2,-1/2}(g(\cdot))(t);$$

en particulier $\check g$ est continue donc $\check g\in\mathcal{C}_{\delta,a}$ et par la proposition 4.5(i) $\mathcal{L}_{\alpha,\beta}(\check g)=g$.

Cas particulier: $\beta=-\frac12$. En tenant compte des remarques 4.6(ii) et 5.2(ii), la théorème ci-dessus peut se reformuler sous la forme suivante dans ce cas limite.

COROLLAIRE 5.5. *Soit $\alpha\in\mathbb{C}$, $\mathrm{Re}(\alpha)>\beta=-\frac12$, soit $a>0$ et $\delta>0$. Pour $f\in\mathcal{C}^\infty_{\delta,a}$, $\mathcal{L}_{\alpha,-1/2}(f)\in\mathcal{H}_{\delta,a-\mathrm{Re}(\rho)+\varepsilon}$ pour tout $\varepsilon>0$ et pour $b>a-\mathrm{Re}(\rho)$*

$$f(t)=\frac1\pi\int_{ib-\infty}^{ib+\infty}\mathcal{L}_{\alpha,-1/2}(f)(\lambda)\psi_{\alpha,-1/2}(\lambda,t)\,d\lambda.$$

Remarques 5.6. (i) Un corollaire analogue peut être obtenu lorsque $\alpha=\beta$ en utilisant les formules de passages (7) et la remarque 4.6(iii).

(ii) Soit $-\mathrm{Re}(\rho)<a<0$, pour $\mathrm{Re}(\alpha)>\mathrm{Re}(\beta)>-\frac12$, pour $g\in\mathcal{H}_{\delta,a}$ et g paire, on a:

$$\check g(t)=\frac1\pi\int_{-\infty}^{+\infty}g(\lambda)\psi_{\alpha,\beta}(\lambda,t)\,d\lambda=\frac{2^{2\rho+1}}{\Gamma(\alpha+1)\sqrt\pi}\int_0^\infty g(\lambda)\omega_{\alpha,\beta}(t)\varphi_{\alpha,\beta}(\lambda,t)\,d\lambda.$$

(iii) Soit $\alpha,\beta\in\mathbb{C}$, $\mathrm{Re}(\alpha)>\mathrm{Re}(\beta)>-\frac12$, soit $0<a<\mathrm{Re}(\rho)$ et soit $f\in\mathcal{C}^\infty_{\delta,a}$, les deux formules d'inversion sont valables et on a donc:

$$f(t)=\frac1\pi\int_{-\infty}^{+\infty}\mathcal{L}_{\alpha,\beta}(f)(\lambda)\psi_{\alpha,\beta}(\lambda,t)\,d\lambda=\frac{2^{2\rho}}{\Gamma(\alpha+1)\sqrt\pi}\int_{-\infty}^{+\infty}\mathcal{L}_{\alpha,\beta}(f\omega_{\alpha,\beta})(\lambda)\varphi_{\alpha,\beta}(\lambda,t)\,d\lambda.$$

Si de plus, $\alpha,\beta\in\mathbb{R}$ et f à valeurs réelles, on a $\mathcal{L}_{\alpha,\beta}(f)(-\lambda)=\overline{\mathcal{L}_{\alpha,\beta}(f)(\lambda)}$ et donc

$$f(t)=\frac{2^{2\rho}}{\Gamma(\alpha+1)\sqrt\pi}\int_0^\infty 2\mathrm{Re}\left[\mathcal{L}_{\alpha,\beta}(f\omega_{\alpha,\beta})(\lambda)\right]\varphi_{\alpha,\beta}(\lambda,t)\,d\lambda;$$

on retrouve la formule [1, (3.13), théorème 3.5], de R. Carroll.

6. Transformations de Bessel; opérateurs de Chebli. Une situation analogue à la transformation de Laplace–Jacobi est fournie par une transformation de Bessel.

Soit la mesure $A_\nu(t)\,dt = t^{2\nu+1}\,dt$ sur $\mathbb{R}_+$; où $\nu \in \mathbb{C}$.

Soit $\Delta_\nu = (1/A_\nu(t))\,\frac{d}{dt}(A_\nu(t)\,\frac{d}{dt})$; dans ce cas $\rho = \lim_{t\to\infty}(A'_\nu(t)/A_\nu(t)) = 0$.

Considérons $\varphi_\nu(\lambda,t)$ et $\Phi_\nu(\lambda,t)$ les solutions de l'opérateur de Bessel: $\Delta_\nu = -\lambda^2$ pour $\lambda \in \mathbb{C}$; ces solutions vérifient:

$$\varphi_\nu(\lambda,0) = 1, \qquad \varphi'_\nu(\lambda,0) = 0,$$
$$\Phi_\nu(\lambda,t)t^{\nu+1/2} = e^{i\lambda t}(1+\varepsilon(t)) \quad \text{avec } \varepsilon(t)\to 0 \quad \text{quand } t\to\infty.$$

On a $\varphi_\nu(\lambda,t) = \Gamma(\nu+1)(\lambda t/2)^{-\nu}J_\nu(\lambda t)$ et $\Phi_\nu(\lambda,t) = ((-i\lambda)^{1/2}/\pi)(K_\nu(-i\lambda t)/t^\nu)$ où J_ν est la fonction de Bessel de 1-ère espèce et K_ν une autre fonction de Bessel (parfois appelée fonction de Bessel de 3-ème espèce).

Soit la transformation que nous appellerons transformation de Laplace–Bessel: $\mathcal{L}_\nu(f)(\lambda) = \int_0^\infty f(t)\Phi_\nu(\lambda,t)t^{2\nu+1}\,dt$. Alors en utilisant les propriétés de la transformation de Meijer, cf. Ditkine et Proudnikov [3, Chap. III], on obtient la transformation réciproque:

$$\mathcal{L}_\nu^{-1}(g)(t) = \int_{ic-\infty}^{ic+\infty} g(\lambda)\psi_\nu(\lambda,t)\lambda^{\nu+1/2}\,d\lambda,$$

où $\psi_\nu(\lambda,t) = (-i)^{1/2}(-i\lambda t)^{-\nu}J_\nu(\lambda t)$ est proportionnelle à $\varphi_\nu(\lambda,t)$.

Il nous semble que le contexte général pour unifier ces transformations de Laplace généralisées soit celui des opérateurs de Chébli:

Soit $A(t)\,dt$ une mesure sur $\mathbb{R}_+$ définie par une fonction A vérifiant les hypothèses suivantes (cf. H. Chébli [2]):

1°) $\alpha \in \mathbb{R}$, $A'(t)/A(t) = \alpha/t + B(t)$ où B est une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ et impaire.

2°) $A(0) = 0$, A croissante et A tend vers l'infini avec t; A'/A décroisante.

Posons $\rho = \frac{1}{2}\lim_{t\to\infty}(A'(t)/A(t))$; soit Δ_A l'opérateur $(1/A(t))\,\frac{d}{dt}(A(t)\,\frac{d}{dt})$; pour $\lambda \in \mathbb{C}$; il existe une solution $\varphi_A(\lambda,t)$ de $\Delta_A f = -(\lambda^2+\rho^2)f$ telle que $\varphi_A(\lambda,0) = 1$ et $\varphi'_A(\lambda,0) = 0$. De plus dans les cas importants que nous avons vu, il existe une deuxième solution linéairement indépendante de φ_A telle que

$$\Phi_A(\lambda,t)\sqrt{A(t)} = e^{i\lambda t}(1+\varepsilon(t)) \qquad (\text{avec } \varepsilon(t)\to 0 \quad \text{quand } t\to\infty).$$

Dans ce cadre, la A-transformation de Fourier $f \to \mathcal{F}_A(f)$ est bien définie par

$$\mathcal{F}_A(f)(\lambda) = \int_0^\infty f(t)\varphi_A(\lambda,t)A(t)\,dt.$$

Une analyse harmonique a été établie (théorème de Plancherel, formule d'inversion, théorème de Paley-Wiener, etc.) par H. Chébli [2] et plus récemment développée par K. Trimèche [13].

On peut alors définir une A-transformation de Laplace en posant $\mathcal{L}_A(f)(\lambda) = \int_0^\infty f(t)\Phi_A(\lambda,t)A(t)\,dt$, ou encore, en tenant compte du coefficient de normalisation $c_A(\lambda)$ défini à partir du Wronskien W de $\Phi_A(\lambda,\cdot)$ et $\varphi_A(\lambda,\cdot)$ par la formule

$$-2i\lambda c_A(\lambda) = A(t)W[\Phi_A(-\lambda,t),\varphi_A(\lambda,t)],$$

on peut définir la A-transformation de Laplace en posant par exemple:

$$\mathcal{L}_A(f)(\lambda) = \int_0^\infty f(t)\,\frac{\Phi_A(\lambda,t)}{c_A(\lambda)}\,dt. \tag{35}$$

Pour avancer dans l'étude d'une A-transformation de Laplace, on pourra partir des transformations intégrales de Riemann-Liouville et de Weyl généralisées associées à la fonction $A(t)$ par K. Trimèche [13]. Ces transformations sont liées à l'opérateur de transmutation $\mathfrak{X}$ tel que: $(\Delta_A+\rho^2)\mathfrak{X}f=\mathfrak{X}(d^2/dt^2)f$ (voir également J. L. Lions [10]).

7. Transformations radiales sur certains groupes de Lie semi-simples. Le corps $\mathbb{K}$ désigne soit le corps $\mathbb{R}$, soit le corps $\mathbb{C}$, soit le corps $\mathbb{H}$ des quaternions. Soit $d=\dim_{\mathbb{R}}(\mathbb{K})$ et soit p et q deux entiers strictement positifs. Soit $\mathbb{K}^{p+q}$ muni de la pseudo-métrique (p,q) associée à la pseudo-norme définie pour $x=(x_i)_{i=1,\dots,p+q}$ par $\|x\|^2_{p,q}=\sum_{i=1}^{p}|x_i|^2 - \sum_{i=p+1}^{q}|x_i|^2$. Considérons les groupes de Lie connexes semi-simples et leurs sous-groupes suivants qui agissent canoniquement sur $\mathbb{K}^{p+q}$ en laissant la pseudo-métrique (p,q) invariante:

$d=1$	$d=2$	$d=4$
$G=SO_0(p,q)$	$G=SU(p,q)$	$G=\mathrm{Sp}(p,q)$
$K=SO(p)\times SO(q)$	$K=S(U(p)\times U(q))$	$K=\mathrm{Sp}(p)\times\mathrm{Sp}(q)$
$H=SO_0(p-1,q)$	$H=S(U(p-1,q)\times U(1))$	$H=\mathrm{Sp}(p-1,q)\times\mathrm{Sp}(1)$

Soit $G=KAN$ la décomposition d'Iwasawa de G relative au sous-groupe compact maximal K. Il existe un sous-groupe à un paramètre A_0 de A tel que nous avons une décomposition du type Iwasawa et une décomposition du type Cartan relatives au sous-groupe H. Plus précisément, soit $W_{H,K}$ le groupe de Weyl défini par $W_{H,K}=M^*_{H\cap K}(A_0)/M_{H\cap K}(A_0)$ où $M^*_{H\cap K}(A_0)$ est le normalisateur de A_0 dans $H\cap K$ et $M_{H\cap K}(A_0)$ est le centralisateur de A_0 dans $H\cap K$.

PROPOSITION 7.1. (Oshima et Sékiguchi [11]). i) *Décomposition du type Iwasawa: il existe un sous-groupe nilpotent N_0 de G tel que $G=\overline{HA_0N_0}$ et tel que si $g\in HA_0N_0$, la décomposition est unique.*

ii) *Décomposition du type Cartan: $G=KA_0H$ ou plus précisément $K/M_{H\cap K}(A_0)\times(A_0-\{e\})/W_{H,K}$ s'identifie à un ouvert partout dense de l'espace pseudo-riemannien symétrique G/H.*

iii) *Une mesure de Haar dg sur G s'écrit: $dg=dkA_{p,q,d}(t)\,dt\,dh$ sur la décomposition $G=KA_0H$ où dk et dh sont des mesures de Haar sur K et H respectivement et où $A_{p,q,d}(t)=(\mathrm{sh}\,t)^{dq-1}(\mathrm{ch}\,t)^{dp-1}$.*

iv) *La restriction de l'opérateur de Casimir Ω à l'ensemble des fonctions analytiques sur G, invariantes à droite par H et à gauche par K est donnée par:*

$$[2d(p+q+2)-8]\Omega=\frac{d^2}{dt^2}+[(dq-1)\coth t+(dp-1)\,\mathrm{th}\,t]\,\frac{d}{dt}.$$

Cette proposition (très classique pour $p=1$) résume des cas particuliers de résultats établis par Oshima et Sékiguchi [11] lorsque $q=1$ et par Sékiguchi [12] lorsque $p>1$ et $q>1$. Ces auteurs, en utilisant également une décomposition du type décomposition de Bruhat à partir du sous-groupe M_0 centralisateur de A_0 dans H, étudient une frontière $\Gamma=G/M_0A_0N_0$ de l'espace pseudo-riemannien symétrique G/H et définissent une transformation de Poisson pour laquelle ils obtiennent des résultats analogues à ceux (classiques) correspondant aux espaces riemanniens symétriques. Voir également Faraut [7].

Ainsi, l'espace des doubles classes $K\backslash G/H$ s'identifie à $\mathbb{R}_+$ (parfois à $\mathbb{R}$ lorsque $q=d=1$), muni de la mesure $A_{p,q,d}(t)\,dt$ et du Laplacien $(1/A_{p,q,d}(t))\frac{d}{dt}(A_{p,q,d}(t)\frac{d}{dt})$. Nous avons donc une transformation de Fourier $\mathfrak{F}_{(dq-2)/2,(dp-2)/2}$ et une transformation de Laplace $\mathfrak{L}_{(dq-2)/2,(dp-2)/2}$ pour les fonctions sur G invariantes à droite par H et

à gauche par K. De plus, lorsque $dp \leq dq+2$ la formule (10) précise la transformation inverse de la transformation de Fourier, et lorsque $p \leq q$ le théorème 5.4 précise la transformation inverse de la transformation de Laplace.

8. Eléments pour une interprétation géométrique sur les groupes $G=SO_0(1,n)$. Dans ce paragraphe, nous examinons le cas particulier $\beta=-\frac{1}{2}$, $\mathrm{Re}(\alpha)>-\frac{1}{2}$. En effet pour $\alpha=(n-2)/2$, où n est un entier supérieur ou égal à 2, l'interprétation géométrique passe par le groupe $SO_0(1,n)$.

RAPPELS. Soit $G=SO_0(1,n)$, soit $G=KAN$ sa décomposition d'Iwasawa, KA^+K, celle de Cartan. Soit M le centralisateur de A dans K, $K \simeq SO(n)$; $A \simeq \mathbb{R}$, $M \simeq SO(n-1)$; soit $H \simeq SO_0(1,n-1)$ le sous-groupe de G admettant M comme sous-groupe compact maximal.

Les fonctions sphériques $\varphi_{\alpha,-1/2}(\lambda,t)$ sont les moyennes d'une puissance du noyau de Poisson, moyenne sur le bord K/M de l'espace homogène G/K. Le noyau de Poisson $P(g,\gamma)$ défini sur $G\times K/M$ étant la dérivée de Radon-Nikodým de l'action canonique de G sur $K/M=G/MAN$. Plus précisément, soit $g\in G$ et $\gamma=\dot{g}_1\in\Gamma=K/M=G/MAN$, alors $g\cdot\gamma=\widehat{gg_1}\in\Gamma$; soit $d\gamma$ la mesure image sur K/M de la mesure de Haar dk sur K, alors $P(g,\gamma)=(dg\cdot\gamma)/d\gamma$ c'est-à-dire pour toute fonction continue f à support compact sur Γ, $\int_\Gamma f(g\cdot\gamma)P(g,\gamma)\,d\gamma=\int_\Gamma f(\gamma)\,d\gamma$, d'où la formule de 2-cocycle vérifiée par le noyau de Poisson $P(g_1g_2,\gamma)=P(g_1,g_2\cdot\gamma)P(g_2,\gamma)$ pour $g_1,g_2\in G$ et $\gamma\in\Gamma$. Cela s'exprime par la formule:

$$\varphi_{\alpha,-1/2}(\lambda,t)=\frac{\Gamma(\alpha+1)}{\sqrt{\pi}\,\Gamma(\alpha+\frac{1}{2})}\int_0^\pi \frac{(\sin\theta)^{2\alpha}d\theta}{(\mathrm{ch}\,t+\mathrm{sh}\,t\cos\theta)^{1/2+\alpha-i\lambda}}, \tag{36}$$

formule valable pour $\mathrm{Re}(\alpha)>-\frac{1}{2}$. Cette formule s'obtient directement en faisant le changement de variable $e^s=\mathrm{ch}\,t+\mathrm{sh}\,t\cos\theta$ dans la formule:

$$2^{2\alpha+1}\omega_{\alpha,-1/2}(t)\varphi_{\alpha,-1/2}(\lambda,t)=\frac{\Gamma(\alpha+1)}{\sqrt{\pi}}\int_0^t \cos\lambda s\,A_{\alpha,-1/2}(s,t)\,ds$$

où

$$A_{\alpha,-1/2}(s,t)=\frac{2^{3(\alpha+1/2)}\,\mathrm{sh}\,t}{\Gamma(\alpha+\frac{1}{2})(\mathrm{ch}\,t-\mathrm{ch}\,s)^{1/2-\alpha}};$$

cf. formule (16)(ii), (iii). Dans ces formules, t paramètre le sous-groupe A de $SO_0(1,n)$.

Peut-on trouver une formule similaire à (36) pour les fonction $\Phi_{\alpha,-1/2}(\lambda,t)$?

PROPOSITION 8.1. *Pour* $\mathrm{Re}(\alpha)>-\frac{1}{2}$ *et pour* $\mathrm{Im}(\lambda)>\mathrm{Re}(\alpha)-\frac{1}{2}$,

$$\Phi_{\alpha,-1/2}(\lambda,t)=\frac{2^{-2\alpha}\Gamma(1-i\lambda)}{\Gamma(\alpha+\frac{1}{2})\Gamma(\frac{1}{2}-\alpha-i\lambda)}\int_0^{+\infty}\frac{(\mathrm{sh}\,\psi)^{2\alpha}d\psi}{(\mathrm{ch}\,t+\mathrm{sh}\,t\,\mathrm{ch}\,\psi)^{1/2+\alpha-i\lambda}}. \tag{37}$$

Cette formule (37) se trouve dans l'article de L. Durand [4, formule (16)]. Cette formule nous suggère de la réaliser comme moyenne "d'un noyau de Poisson hyperbolique" sur un "bord" de l'hyperboloïde $SO_0(1,n)/SO(n)$ réalisé dans $\mathbb{R}^{n+1}$; c'est ce que nous allons faire:

Quelques notations. Soit $G=SO_0(1,n)$, soit $K=SO(n)$ le sous-groupe compact maximal de G et $M=SO(n-1)$ le centralisateur de A dans K, où A est le sous-groupe de G provenant de la décomposition d'Iwasawa de $G=KAN$. Soit $H=SO_0(1,n-1)$ le sous-groupe non compact de G admettant M comme sous-groupe compact maximal. Soit $P=MAN$ le parabolique minimal de G.

Nous noterons encore $X=G/K$ l'espace riemannien symétrique et $\Gamma=G/P=K/M$ sa frontière; nous noterons encore $\Gamma_1=H/M$ une frontière de l'espace affine symétrique G/H; pour plus de précisions on pourra se reporter à [11] dont les propositions 1.10 et 2.7 permettent de dire que $H\times A\times N\to HAN$ est un diffiomorphisme de $H\times A\times N$ sur un ouvert de G et que Γ_1 peut s'identifier à un ouvert de Γ. Plus précisément:

Soit

$$A=\left\{a_t=\begin{pmatrix} \operatorname{ch} t & \operatorname{sh} t & & \\ \operatorname{sh} t & \operatorname{ch} t & & \\ & & I_{n-1} & \end{pmatrix} / t\in\mathbb{R}\right\}, \qquad A^+=\{a_t\in A/t\in\mathbb{R}_+\}.$$

Soit

$$A_K=\left\{k_\theta=\begin{pmatrix} 1 & & & \\ & \cos\theta & -\sin\theta & \\ & \sin\theta & \cos\theta & \\ & & & I_{n-2} \end{pmatrix} / \theta\in[0,2\pi[\right\}, \qquad A_K^+=\{k_\theta\in A/\theta\in[0,\pi[\}.$$

Soit

$$A_H=\left\{h_\psi=\begin{pmatrix} \operatorname{ch}\psi & 0 & \operatorname{sh}\psi & \\ 0 & 1 & 0 & \\ \operatorname{sh}\psi & 0 & \operatorname{ch}\psi & \\ & & & I_{n-2} \end{pmatrix} / \psi\in\mathbb{R}\right\}, \qquad A_H^+=\{h_\psi\in A_H/\psi\in\mathbb{R}_+\}.$$

Soit

$$M=\left\{m=\begin{pmatrix} 1 & & \\ & 1 & \\ & & \tilde{m} \end{pmatrix} / \tilde{m}\in SO(n-1)\right\}, \qquad W=\{I_{n+1}, k_\pi\}$$

Nous utiliserons les décompositions de Cartan de K et H: $K=MA_K^+M$ et $H=MA_H^+M$ munis des mesures $dk=dm(\sin\theta)^{n-2}\,d\theta\,dm$ et $dh=dm(\operatorname{sh}\psi)^{n-2}\,d\psi\,dm$, où dm est la mesure de Haar normalisée sur M. Le groupe G agit canoniquement sur l'espace de Minkowski $\mathbb{R}^{n+1}$. Soit C le cône de $\mathbb{R}^{n+1}$ et Ξ l'ensemble des génératrices de ce cône C. Par l'application

$$k\to k\begin{pmatrix} 1 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

Γ s'identifie à Ξ et par

$$g\to g\begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

l'espace symétrique $G/K\simeq K/M\times A^+$ s'identifie à la nappe supérieure de l'hyperboloïde à deux nappes de $\mathbb{R}^{n+1}$.

D'autre part G agit canoniquement sur Γ; soit $P(g,\gamma)$ la dérivée de Radon–Nikodym de cette action sur $\Gamma \simeq M/M' \times A_K^+$ muni de la mesure $d\dot{m}\,d\theta$, où M' est le centralisateur de A_K dans M et $d\dot{m}$ la mesure canonique sur M/M'. On a (cf. [6] par exemple)

$$\varphi_{(n-2)/2,-1/2}(\lambda,t)=\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\int_\Gamma P(a_t,\gamma)^{(n-1)/2-i\lambda}d\gamma$$

où $d\gamma = d\dot{m}(\sin\theta)^{n-2}d\theta$ est la mesure canonique sur K/M. De plus par l'identification entre Γ et Ξ nous retrouvons exactement la formule (36).

Interprétation gémétrique de la formule (37). Soit

$$\Xi_1 = H\begin{pmatrix}1\\1\\0\\\vdots\\0\end{pmatrix};$$

alors la frontière $\Gamma_1 = H/M \simeq M/M' \times A_H^+$ s'identifie à Ξ_1, partie ouverte de Ξ; soit

$$\Xi_2 = Hk_\pi\begin{pmatrix}1\\1\\0\\\vdots\\0\end{pmatrix},$$

c'est la deuxième frontière de l'espace affine symétrique G/H que l'on peut identifier à l'hyperboloïde à une nappe de $\mathbb{R}^{n+1}$. Les espaces Ξ_1 et Ξ_2 sont deux ouverts disjoints de Ξ et leur réunion est partout dense dans Ξ, cf. [11].

G agit sur Ξ; considérons G_1 l'ensemble des éléments de G laissant Ξ_1 stable; on obtient:

LEMME 8.2. (i) *Pour a_{t_1} et $a_{t_2} \in A^+$ et pour $h_\psi \in H$, on a $a_{t_1}h_\psi a_{t_2} = h_{\psi_1}a_t h_{\psi_2}$, où h_{ψ_1} et $h_{\psi_2} \in H$ et où a_t unique dans A^+ est défini par* $\operatorname{ch} t = \operatorname{ch} t_1 \operatorname{ch} t_2 + \operatorname{sh} t_1 \operatorname{sh} t_2 \operatorname{ch}\psi$.

(ii) *G_1 est un sous-semi-groupe de G, égal à HA^+H. Plus précisément, pour tout élément $g \in G_1$, il existe h_{ψ_1} et $h_{\psi_2} \in H$, et un unique élément $a_t \in A^+$ tels que $g = h_{\psi_1}a_t h_{\psi_2}$.*

(i) Provient d'un calcul matriciel évident et (ii) est une conséquence de (i) et du fait que G_1 est le sous-semi-groupe de G engendré par H et A^+.

Soit Ξ_1 paramétré par $\Gamma_1 = M/M' \times A_H^+$ et muni de la mesure associée $d\gamma_1 = d\dot{m}(sh\psi)^{n-2}d\psi$. Posons $Q(g,\gamma_1)$ la dérivée de Radon–Nikodým de l'action de G_1 sur Ξ_1 muni de la mesure $d\dot{m}\,d\psi$: on a $Q(h,\gamma_1)=1$; calculons $Q(a_t,\gamma_1) = Q(a_t,(\dot{m},h_\psi))$:

$$a_t\cdot(\dot{m},h_\psi)\begin{pmatrix}1\\1\\0\\\vdots\\0\end{pmatrix} = a_t\begin{pmatrix}\operatorname{ch}\psi\\1\\\vdots\end{pmatrix} = \begin{pmatrix}\operatorname{sh} t + \operatorname{ch} t \operatorname{ch}\psi\\ \operatorname{ch} t + \operatorname{sh} t \operatorname{ch}\psi\\ \vdots\end{pmatrix};$$

posons

$$a_t \cdot (\dot{m}, h_\psi) \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = (\dot{m}(t), h_{\psi(t)}) \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

on obtient par un calcul élémentaire:

$$\frac{d\dot{m}(t)\, d\psi(t)}{d\dot{m}\, d\psi} = \frac{d\psi(t)}{d\psi} = \frac{1}{\operatorname{ch} t + \operatorname{sh} t \operatorname{ch}\psi}.$$

Ainsi on peut réécrire la formule (37) de la manière suivante:

PROPOSITION 8.3. *Lorsque* $\operatorname{Im}(\lambda) > \frac{n-3}{2}$ *on a*:

$$(38) \quad \Phi_{(n-2)/2,-1/2}(\lambda, t) = \frac{2^{2-n}\Gamma(1-i\lambda)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{3-n}{2} - i\lambda\right)} \int_{\Gamma_1} Q(a_t, \gamma_1)^{(n-1)/2 - i\lambda} d\gamma_1.$$

C'est-à-dire les fonctions de Jacobi de deuxième espèce sont moyennes du noyau de Poisson hyperbolique $Q(a_t, \gamma_1)$ sur la première frontière $SO_0(1, n-1)/SO(n-1)$ de l'espace riemannien symétrique $SO_0(1,n)/SO(n)$.

Remarque. Soit $\alpha \in \mathbb{C}$, $\operatorname{Re}(\alpha) > -\frac{1}{2}$, définissons une action de $t \in \mathbb{R}_+$ sur $\psi \in \mathbb{R}^+$ définie par $t \cdot \psi = \psi(t)$ tel que $\operatorname{ch}\psi(t) = (\operatorname{sh} t + \operatorname{ch} t \operatorname{ch}\psi)/(\operatorname{ch} t + \operatorname{sh} t \operatorname{ch}\psi)$ on a $Q(t,\psi) = d\psi(t)/d\psi = (\operatorname{ch} t + \operatorname{sh} t \operatorname{ch}\psi)^{-1}$. Ainsi $\Phi_{\alpha,-1/2}(\lambda, t)$ apparaît comme moyenne sur $\mathbb{R}_+$ muni de la mesure $(\operatorname{sh}\psi)^{2\alpha} d\psi$ du noyau de Poisson hyperbolique $Q(t,\psi)$.

Formule d'addition des fonctions de Jacobi de 2-ème espèce. Considérons Φ comme une fonction définie sur G_1, bi-invariante par H, en posant:

$$\Phi_{(n-2)/2,-1/2}(\lambda, h_1 a_t h_2) = \Phi_{(n-2)/2,-1/2}(\lambda, t).$$

En utilisant le lemme 8.2 et la proposition 8.3, on obtient par des arguments classiques la formule d'addition suivante (cf., mutatis mutandis, la démonstration dans P. Eymard [6, théorème 4]).

PROPOSITION 8.4. i) *Pour tout* $g_1, g_2 \in G_1$ *et pour tout* $\gamma \in \Gamma_1$:

$$Q(g_2 g_1, \gamma) = Q(g_2, g_1 \cdot \gamma) Q(g_1, \gamma);$$

ii) *pour tout* $g_1, g_2 \in G_1$ *et pour* $\operatorname{Im}(\lambda) > \frac{n-3}{2}$:

$$(39) \quad \Phi_{(n-2)/2,-1/2}(\lambda, g_1)\Phi_{(n-2)/2,-1/2}(\lambda, g_2) = \frac{2^{2-n}\Gamma(1-i\lambda)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{3-n}{2} - i\lambda\right)} \int_H \Phi_{(n-2)/2,-1/2}(\lambda, g_1 h g_2)\, dh.$$

On retrouve un cas particulier important d'un résultat récent de L. Durand, cf. [4], que l'on peut écrire comme suit: Pour $\operatorname{Re}(\alpha) > -\frac{1}{2}$ et $\operatorname{Im}(\lambda) > \operatorname{Re}(\alpha) - \frac{1}{2}$

$$\Phi_{\alpha,-1/2}(\lambda, t_1)\Phi_{\alpha,-1/2}(\lambda, t_2) = \frac{2^{-2\alpha}\Gamma(1-i\lambda)}{\Gamma(\alpha + \frac{1}{2})\Gamma(\frac{1}{2} - \alpha - i\lambda)} \times \int_0^\infty \Phi_{\alpha,-1/2}(\lambda, \operatorname{Arg ch}(\operatorname{ch} t_1 \operatorname{ch} t_2 + \operatorname{sh} t_1 \operatorname{sh} t_2 \operatorname{ch}\psi))(\operatorname{sh}\psi)^{2\alpha} d\psi.$$

Note. M. J. Faraut me signale, entre autres résultats, qu'il vient de munir l'espace des fonctions à support dans G_1, et bi-invariantes par H, d'une structure d'algèbre de convolution commutative telle que les fonctions $\Phi_{\alpha,-1/2}(\lambda, g)$ définissent des caractères de cette algèbre.

REFERENCES

[1] R. CARROLL, *Some inversion theorems of Fourier type*, preprint (1980).

[2] H. CHÉBLI, *Sur un théorème de Paley-Wiener associé à la décomposition spectrale d'un opérateur de Sturm-Liouville sur* $]0, +\infty[$, J. Functional Anal. 17 (1974), pp. 447–461.

[3] V. DITKINE AND A. PROUDNIKOV, *Transformations intégrales et calcul opérationnel*, Editions MIR, Moscow, 1978.

[4] L. DURAND, *Nicholson-type integrals for products of Gegenbauer functions and related topics*, dans Theory and Application of Special Functions, R. Askey, Academic Press, New York, 1975.

[5] A. ERDÉLYI ET AL., *Tables of Integral Transforms*, McGraw-Hill, New York, 1954, vol. 2.

[6] P. EYMARD, *Le noyau de Poisson et la théorie des groupes*, Symposia Math., XXII, INDAM, Rome, 1977, Academic Press, New York, 1977, pp. 107–132.

[7] J. FARAUT, *Distributions sphériques sur les espaces hyperboliques*, J. Math. Pures Appl., 58 (1979), pp. 369–444.

[8] S. HELGASON, *Functions on symmetric spaces*, Proc. of Symposia in Pure Math., 26, 1972, American Mathematical Society, Providence, RI, 1973, pp. 102–146.

[9] T. KOORNWINDER, *A new proof of a Paley-Wiener type theorem for the Jacobi transform*, Ark. Mat., 13 (1975), pp. 145–159.

[10] J. L. LIONS, *Opérateurs de Delsarte et problèmes mixtes*, Bull. Soc. Math. France, 84 (1956), pp. 9–95.

[11] T. OSHIMA AND J. SEKIGUCHI, *Eigenspaces of invariant differential operators on an affine symmetric space*, Inventiones Math., 57 (1980), pp. 1–81.

[12] J. SEKIGUCHI, *Eigenspaces of the Laplace–Beltrami operator on a hyperboloid*, Nagoya Math. J., 79 (1980), pp. 151–185.

[13] K. TRIMECHE, *Transformation intégrale de Weyl et théorème de Paley-Wiener associés à un opérateur différentiel singulier sur* $(0, \infty)$, J. Math. Pures Appl., 60 (1981), pp. 51–98.

[14] G. A. VIANO, *On the harmonic analysis of the elastic scattering amplitude of two spinless particles at fixed momentum transfer*, Ann. Inst. H. Poincaré Sect. A, 32 (1980), pp. 109–123.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0014 $01.25/0

AN INEQUALITY OF THE MARKOV–BERNSTEIN TYPE FOR POLYNOMIALS*

L. MIRSKY†

To Professor Alexander Ostrowski on his 90th birthday

Abstract. Let $-\infty \le a < b \le \infty$ and denote by $w:(a,b)\to\mathbb{R}$ a positive and integrable function, with all moments

$$\int_a^b t^n w(t)\,dt$$

finite. For any polynomial f with complex coefficients, we write

$$\|f\| = \left\{\int_a^b |f(t)|^2 w(t)\,dt\right\}^{1/2}.$$

Then there exists a constant γ_n (depending on a, b, w but not on f) such that, for every polynomial f with complex coefficients and of degree $\le n$,

$$\|f'\| \le \gamma_n \|f\|.$$

An admissible value for γ_n can be expressed very simply in terms of the system of orthonormal polynomials associated with the interval (a,b) and the function w.

1. Let f be a polynomial of degree n with complex coefficients, and write

$$\|f\| = \left\{\int_0^1 |f(t)|^2 dt\right\}^{1/2}.$$

E. Schmidt [3] stated, without proof, the inequality

$$\|f'\| \le \frac{(n+1)^2}{\sqrt{2}}\|f\|, \tag{1}$$

and R. Bellman [1] established this result in a remarkably simple and elegant way by using elementary identities for derivatives of Legendre polynomials. It is not difficult, by pursuing Bellman's method a little further, to improve the constant $(n+1)^2/\sqrt{2}$ to $n^2/2\sqrt{2} + O(n)$. However, there is little point in this elaboration since, by a more sophisticated method involving properties of ultraspherical polynomials, Hille, Szegö, and Tamarkin [2] had previously strengthened (1) in a much more decisive fashion.

In the present communication we follow in the footsteps of Bellman by establishing a more general result of the same type as (1).

2. Let $-\infty \le a < b \le \infty$; let $w:(a,b)\to\mathbb{R}$ be positive and integrable; and suppose that all moments of w, namely

$$\int_a^b t^n w(t)\,dt \qquad (n \ge 0),$$

* Received by the editors October 20, 1982.
† Department of Pure Mathematics, University of Sheffield, Sheffield, S3 7RH England.

are finite. We recall the existence of a sequence $(p_n: n\geq 0)$ of (real) polynomials such that $\deg p_n = n$ $(n\geq 0)$ and

$$\int_a^b p_m(t)p_n(t)w(t)\,dt=\delta_{mn} \qquad (m,n\geq 0). \tag{2}$$

(These polynomials are unique to within factors ± 1.) It follows, in particular, that

$$\int_a^b t^k p_n(t)w(t)\,dt=0 \qquad (0\leq k<n). \tag{3}$$

For any polynomial f (with complex coefficients) we shall write

$$\|f\|=\left\{\int_a^b |f(t)|^2 w(t)\,dt\right\}^{1/2}.$$

PROPOSITION. *There exists a number $\gamma_n=\gamma_n(a,b;w)$ such that, for every polynomial f with complex coefficients and of degree not exceeding n, we have*

$$\|f'\|\leq\gamma_n\|f\|. \tag{4}$$

A routine verification shows that it suffices to establish (4) for polynomials with real coefficients. We may clearly write

$$f(t)=\sum_{0\leq k\leq n} c_k p_k(t), \qquad f'(t)=\sum_{0\leq j\leq n-1} d_j p_j(t),$$

where the c's and d's are uniquely determined. Hence, by (2),

$$\|f\|^2=\sum_{0\leq k\leq n} c_k^2, \qquad \|f'\|^2=\sum_{0\leq j\leq n-1} d_j^2.$$

Next, put

$$e_{kj}=\int_a^b p_k'(t)p_j(t)w(t)\,dt$$

so that plainly, by (3),

$$e_{kj}=0 \qquad (k\leq j). \tag{5}$$

Now we have

$$\sum_{0\leq k\leq n} c_k p_k'(t)=\sum_{0\leq r\leq n-1} d_r p_r(t) \qquad (=f'(t)).$$

Hence, for $0\leq j\leq n-1$,

$$\sum_{0\leq k\leq n} c_k p_k'(t)p_j(t)w(t)=\sum_{0\leq r\leq n-1} d_r p_r(t)p_j(t)w(t)$$

and so

$$\sum_{0\leq k\leq n} c_k e_{kj}=d_j.$$

Thus, by (5),

$$d_j=\sum_{j+1\leq k\leq n} c_k e_{kj}.$$

It follows that

$$\begin{aligned} d_j^2 &\le \Big(\sum_{j+1\le k\le n} c_k^2\Big)\Big(\sum_{j+1\le k\le n} e_{kj}^2\Big) \\ &\le \Big(\sum_{0\le k\le n} c_k^2\Big)\Big(\sum_{j+1\le k\le n} e_{kj}^2\Big) \\ &= \|f\|^2 \sum_{j+1\le k\le n} e_{kj}^2. \end{aligned}$$

The proof of (4) can now be completed at once by observing that

$$\|f'\|^2 = \sum_{0\le j\le n-1} d_j^2 \le \|f\|^2 \sum_{0\le j\le n-1} \sum_{j+1\le k\le n} e_{kj}^2 \tag{6}$$

and that the double sum on the right-hand side of (6) is independent of f. However, by a slight elaboration we can obtain a very simple formula for an admissible value of γ_n in (4).

We have

$$\begin{aligned} e_{kj}^2 &= \Big\{\int_a^b p_k'(t)\sqrt{w(t)}\cdot p_j(t)\sqrt{w(t)}\,dt\Big\}^2 \\ &\le \int_a^b (p_k'(t))^2 w(t)\,dt \int_a^b p_j^2(t)w(t)\,dt = \|p_k'\|^2. \end{aligned}$$

Hence, by (6),

$$\begin{aligned} \|f'\|^2 &\le \|f\|^2 \sum_{0\le j\le n-1} \sum_{j+1\le k\le n} \|p_k'\|^2 \\ &= \|f\|^2 \sum_{1\le k\le n} \|p_k'\|^2 \sum_{0\le j\le k-1} 1 = \|f\|^2 \sum_{1\le k\le n} k\|p_k'\|^2. \end{aligned}$$

This completes the proof of the proposition, with the value of γ_n given by

$$\gamma_n = \Big\{\sum_{k=1}^n k\|p_k'\|^2\Big\}^{1/2}. \tag{7}$$

3. Let us denote by Γ_n the least admissible value of γ_n in (4). To put it another way, let

$$\Gamma_n = \sup_f \|f'\|/\|f\|,$$

where the supremum is taken with respect to all (nonnull) polynomials f with complex coefficients and of degree $\le n$. By the discussion in §2, we know that

$$\Gamma_n \le \Big\{\sum_{k=1}^n k\|p_k'\|^2\Big\}^{1/2}. \tag{8}$$

The main interest of this result is, however, qualitative, for the bound specified by (8) can be very crude. Consequently, if an accurate estimate of γ_n is needed it is better to pursue an ad hoc approach rather than rely on (8). We shall illustrate this remark by

considering the case $a=-\infty$, $b=\infty$, $w(t)=e^{-t^2}$, which gives rise to the orthogonal system of Hermite polynomials. We shall now write

$$\|f\|^2=\int_{-\infty}^{\infty}|f(t)|^2e^{-t^2}dt.$$

We recall that the Hermite polynomials $(H_n: n\geq 0)$ are defined as coefficients in the expansion

$$e^{2tu-u^2}=\sum_{n=0}^{\infty}\frac{u^n}{n!}H_n(t)$$

and that they satisfy the relations

$$\int_{-\infty}^{\infty}H_m(t)H_n(t)e^{-t^2}dt=\pi^{1/2}2^n n!\delta_{mn} \qquad (m,n\geq 0), \tag{9}$$

$$H_n'(t)=2nH_{n-1}(t) \qquad (n\geq 1). \tag{10}$$

Let f be a polynomial of degree at most n. (As in the earlier discussion, it suffices to consider polynomials with real coefficients.) Write

$$f(t)=\sum_{0\leq k\leq n}\lambda_k H_k(t).$$

Then, by (9),

$$\|f\|^2=\sum_{0\leq k\leq n}\lambda_k^2\pi^{1/2}2^k k!. \tag{11}$$

Again, by (10), (9), and (11),

$$\begin{aligned}
f'(t)&=\sum_{1\leq k\leq n}\lambda_k H_k'(t)=2\sum_{1\leq k\leq n}k\lambda_k H_{k-1}(t),\\
\{f'(t)\}^2&=4\sum_{1\leq k,j\leq n}kj\lambda_k\lambda_j H_{k-1}(t)H_{j-1}(t),\\
\|f'\|^2&=4\sum_{1\leq k\leq n}k^2\lambda_k^2\pi^{1/2}2^{k-1}(k-1)!\\
&=2\sum_{1\leq k\leq n}k\lambda_k^2\pi^{1/2}2^k k!\leq 2n\sum_{1\leq k\leq n}\lambda_k^2\pi^{1/2}2^k k!\\
&=2n\|f\|^2.
\end{aligned}$$

Thus

$$\Gamma_n\leq(2n)^{1/2}.$$

Moreover, for the special choice $f(t)=H_n(t)$, it is easy to verify from (9) and (10) that $\|f'\|=(2n)^{1/2}\|f\|$. Therefore

$$\Gamma_n=(2n)^{1/2}. \tag{12}$$

Let us compare this with the result obtained by the application of the general formula (8). In view of (9), the polynomials p_n are given by

$$p_n(t)=\kappa_n H_n(t), \qquad \kappa_n=(\pi^{1/2}2^n n!)^{-1/2}.$$

Hence

$$p_n'(t)=\kappa_n H_n'(t)=2n\kappa_n H_{n-1}(t)=2n\frac{\kappa_n}{\kappa_{n-1}}p_{n-1}(t)=(2n)^{1/2}p_{n-1}(t),$$

$$\|p_n'(t)\|^2=\int_{-\infty}^{\infty}\{p_n'(t)\}^2e^{-t^2}dt=2n\int_{-\infty}^{\infty}p_{n-1}^2(t)e^{-t^2}dt=2n,$$

$$\Gamma_n\leq\left\{\sum_{k=1}^{n}k\|p_k'\|^2\right\}^{1/2}=\left\{\sum_{k=1}^{n}2k^2\right\}^{1/2},$$

$$\Gamma_n=O(n^{3/2}).$$

The contrast between this estimate and (12) is evident.

REFERENCES

[1] R. BELLMAN, *A note on an inequality of E. Schmidt*, Bull. Amer. Math. Soc., 50 (1944), pp. 734–736.

[2] E. HILLE, G. SZEGÖ AND J. D. TAMARKIN, *On some generalizations of a theorem of A. Markoff*, Duke Math. J., 3 (1937), pp. 729–739.

[3] E. SCHMIDT, *Die asymptotische Bestimmung des Maximums des Integrals über das Quadrat der Ableitung eines normierten Polynoms, dessen Grad ins Unendliche wächst*, Sitzber. Preuss. Akad. Wiss. (1932), p. 287.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0015 $01.25/0

THE ε-ALGORITHM AND PADÉ-APPROXIMANTS IN OPERATOR THEORY*

ANNIE A. M. CUYT†

Abstract. The ε-algorithm of Wynn is closely related to the Padé-table of a univariate function in the following sense: if we apply the ε-algorithm to the partial sums of the power series $f(x)=\sum_{i=0}^{\infty}c_i x^i$ then $\varepsilon_{2m}^{(l-m)}$ is the (l,m) Padé-approximant to $f(x)$ where l is the degree of the numerator and m is the degree of the denominator [C. Brezinski, *Algorithmes d'accélération de la convergence*, Editions Technip, Paris, 1978, pp. 66–68]. In this paper we see that the Padé-approximants for nonlinear operators F: $X\to Y$, where X is a Banach space and Y a commutative Banach algebra, introduced in [Springer Lect. Notes in Math. 765, 1979, pp. 61–87], satisfy the same property as the univariate Padé-approximants.

1. Padé-approximants in operator theory. We briefly repeat the definition of Padé-approximants in operator theory and a determinental formula for their calculation. More details can be found in [3] and [4].

Let X be a Banach space and Y a commutative Banach algebra (0 denotes the unit for the addition and I the unit for the multiplication). Let F: $X\to Y$ be analytic in the open ball $B(0,r)$ with centre $0\in X$ and radius $r>0$ [5, pp. 113]:

$$F(x)=\sum_{k=0}^{\infty}\frac{1}{k!}F^{(k)}(0)x^k \quad \text{for } \|x\|<r,$$

where $F^{(k)}(0)$ is the kth Fréchet-derivative of F in 0 and thus a symmetric k-linear bounded operator, and $(1/0!)F^{(0)}(0)x^0=F(0)$.

DEFINITION 1.1. $F(x)=O(x^k)$ $(k\in\mathbb{N})$ if nonnegative constants $r<1$ and K exist such that $\|F(x)\|\le K\|x\|^k$ for $\|x\|<r$.

Write $D(F)=\{x\in X|F(x)$ is regular in Y, i.e. there exists $y\in Y$: $F(x)\cdot y=I=y\cdot F(x)\}$. We shall denote by y^{-1} the inverse element of y in Y for the multiplication in that Banach algebra.

DEFINITION 1.2. An *abstract polynomial* is a nonlinear operator P: $X\to Y$ with $P(x)=A_n x^n+A_{n-1}x^{n-1}+\cdots+A_0$, where A_i is a symmetric i-linear bounded operator $(i=0,\cdots,n)$ [5, pp. 194].

DEFINITION 1.3. The couple of abstract polynomials

$$(P(x),Q(x))=\left(\sum_{i=0}^{n}A_{nm+i}x^{nm+i},\ \sum_{j=0}^{m}B_{nm+j}x^{nm+j}\right)$$

such that the abstract power series $(F\cdot Q\text{-}P)(x)=O(x^{nm+n+m+1})$ is called a *solution of the Padé-approximation problem of order* (n,m). The choice of $P(x)$ and $Q(x)$, or in other words the translation of degrees in P and Q by $n\cdot m$, can be justified as follows [3]. Write $C_k x^k=(1/k!)F^{(k)}(0)x^k$.

The condition in Definition 1.3 is equivalent with (1a) and (1b):

$$\text{(1a)}\qquad \begin{aligned}
&C_0\cdot B_{nm}x^{nm}=A_{nm}x^{nm} && \forall x\in X,\\
&C_1x\cdot B_{nm}x^{nm}+C_0\cdot B_{nm+1}x^{nm+1}=A_{nm+1}x^{nm+1} && \forall x\in X,\\
&\quad\vdots\\
&C_n x^n\cdot B_{nm}x^{nm}+\cdots+C_0\cdot B_{nm+n}x^{nm+n}=A_{nm+n}x^{nm+n} && \forall x\in X,
\end{aligned}$$

*Received by the editors December 10, 1981, and in revised form June 12, 1982.

†Aspirant N.F.W.O. (Belgium), University of Antwerp, Department of Mathematics, Universiteitsplein 1, B-2610 Wilrijk, Belgium.

with $B_{nm+j}x^{nm+j}\equiv 0$ if $j>m$;

$$C_{n+1}x^{n+1}\cdot B_{nm}x^{nm}+\cdots+C_{n+1-m}x^{n+1-m}\cdot B_{nm+m}x^{nm+m}=0 \quad \forall x\in X,$$

(1b) $\quad\vdots$

$$C_{n+m}x^{n+m}\cdot B_{nm}x^{nm}+\cdots+C_nx^n\cdot B_{nm+m}x^{nm+m}=0 \quad \forall x\in X,$$

with $C_kx^k\equiv 0$ if $k<0$. A solution of (1b) can be computed by means of the following determinants in Y; these formulas are direct generalizations of the classical formulas for the solution of a homogeneous system of m equations in $m+1$ unknowns $B_{nm+j}x^{nm+j}$ $(j=0,\cdots,m)$:

The nm-linear bounded operator

$$\begin{vmatrix} C_nx^n & \cdots & C_{n+1-m}x^{n+1-m} \\ C_{n+1}x^{n+1} & \cdots & C_{n+2-m}x^{n+2-m} \\ \vdots & & \vdots \\ C_{n-1+m}x^{n-1+m} & \cdots & C_nx^n \end{vmatrix} = B_{nm}x^{nm},$$

the $(nm+j)$-linear bounded operator

$$\begin{vmatrix} C_nx^n & \cdots & -C_{n+1}x^{n+1} & \cdots & C_{n+1-m}x^{n+1-m} \\ \vdots & & \vdots & & \vdots \\ C_{n-1+m}x^{n-1+m} & & -C_{n+m}x^{n+m} & \cdots & C_nx^n \end{vmatrix} = B_{nm+j}x^{nm+j},$$

$\uparrow$

jth column in $B_{nm}x^{nm}$ replaced by this column

$1\le j\le m.$

For every solution of (1b) a solution of (1a) can be calculated by substitution of the $B_{nm+j}x^{nm+j}$ $(j=0,\cdots,m)$ in the left hand side of (1a). So using the classical formulas for the solution of a homogeneous system of equations we get immediately the translation of degrees by $n\cdot m$ in $P(x)$ and $Q(x)$. As a result of these formulas we can also write down the following determinental formulas for $P(x)$ and $Q(x)$:

$$Q(x)=\begin{vmatrix} I & \cdots & I \\ C_{n+1}x^{n+1} & C_nx^n & \cdots & C_{n+1-m}x^{n+1-m} \\ \vdots & \vdots & & \vdots \\ C_{n+m}x^{n+m} & C_{n+m-1}x^{n+m-1} & \cdots & C_nx^n \end{vmatrix},$$

$$P(x)=\begin{vmatrix} F_n(x) & F_{n-1}(x) & \cdots & F_{n-m}(x) \\ C_{n+1}x^{n+1} & C_nx^n & \cdots & C_{n+1-m}x^{n+1-m} \\ \vdots & \vdots & & \vdots \\ C_{n+m}x^{n+m} & C_{n+m-1}x^{n+m-1} & \cdots & C_nx^n \end{vmatrix}$$

where $F_i(x)=\sum_{k=0}^{i}C_kx^k$ and $F_i(x)\equiv 0$ for $i<0$. We shall now see how the determinant representations of $P(x)$ and $Q(x)$ link this solution of the Padé-approximation problem of order (n,m) to the ε-algorithm.

2. The ε-algorithm. The ε-algorithm is a nonlinear algorithm due to Wynn [2, pp. 42]; input are the elements of a sequence $\{S_i | i=0,1,\cdots\}$. The following computations are performed:

$$\varepsilon_{-1}^{(i)}=0, \qquad i=0,1,\cdots,$$
$$\varepsilon_{0}^{(i)}=S_i, \qquad i=0,1,\cdots,$$
$$\varepsilon_{2j}^{(-j-1)}=0, \qquad j=0,1,\cdots,$$
$$\varepsilon_{j+1}^{(i)}=\varepsilon_{j-1}^{(i+1)}+\left[\varepsilon_{j}^{(i+1)}-\varepsilon_{j}^{(i)}\right]^{-1}, \qquad j=0,1,\cdots, \quad i=-j,-j+1,\cdots.$$

The $\varepsilon_j^{(i)}$ can be ordered in a table where (i) indicates a diagonal and j a column:

$$\begin{array}{ccccccc}
 & & \varepsilon_0^{(-1)}=0 & & \varepsilon_2^{(-2)}=0 & \cdots \\
\varepsilon_{-1}^{(0)}=0 & & & \varepsilon_1^{(-1)} & & \\
 & & \varepsilon_0^{(0)}=S_0 & & \varepsilon_2^{(-1)} & \cdots \\
\varepsilon_{-1}^{(1)}=0 & & & \varepsilon_1^{(0)} & & \\
 & & \varepsilon_0^{(1)}=S_1 & & \varepsilon_2^{(0)} & \cdots \\
\varepsilon_{-1}^{(2)}=0 & & & \varepsilon_1^{(1)} & & \\
 & & \varepsilon_0^{(2)}=S_2 & & \varepsilon_2^{(1)} & \cdots \\
\varepsilon_{-1}^{(3)}=0 & & \vdots & \varepsilon_1^{(2)} & \vdots & \\
\vdots & & & \vdots & &
\end{array}$$

Let us now take $\{S_i | i=0,1,\cdots\} \subseteq Y$ and denote by $\Delta S_i = S_{i+1}-S_i$ and $\Delta^2 S_i = \Delta S_{i+1}-\Delta S_i$. Write

$$H_j(S_i)=\begin{vmatrix} S_i & \cdots & S_{i+j-1} \\ \vdots & & \vdots \\ S_{i+j-1} & \cdots & S_{i+2j-2} \end{vmatrix}.$$

We can prove the following property for the $\varepsilon_j^{(i)}$. The proof is very technical and similar to the proof in [2, pp. 44–46].

THEOREM 2.1. *If $H_{j-1}(\Delta^2 S_{i+1})$ and $H_j(\Delta^2 S_i)$ are regular in Y, then*

$$\varepsilon_{2j}^{(i)}=\frac{\begin{vmatrix} S_{i+j} & \cdots & & S_i \\ \Delta S_{i+j} & \cdots & \Delta S_{i+1} & \Delta S_i \\ \vdots & \ddots & \vdots & \vdots \\ \Delta S_{i+2j-1} & \cdots & \Delta S_{i+j} & \Delta S_{i+j-1} \end{vmatrix}}{\begin{vmatrix} I & \cdots & I \\ \Delta S_{i+j} & \cdots & \Delta S_i \\ \vdots & & \vdots \\ \Delta S_{i+2j-1} & \cdots & \Delta S_{i+j-1} \end{vmatrix}},$$

and if $H_j(\Delta S_{i+1})$ *and* $H_{j+1}(\Delta S_i)$ *are regular in* Y, *then*

$$\varepsilon_{2j+1}^{(i)} = \frac{\begin{vmatrix} I & \cdots & I \\ \Delta^2 S_{i+j} & \cdots & \Delta^2 S_i \\ \vdots & & \vdots \\ \Delta^2 S_{i+2j-1} & \cdots & \Delta^2 S_{i+j-1} \end{vmatrix}}{\begin{vmatrix} \Delta S_{i+j} & \cdots & \Delta S_i \\ \Delta^2 S_{i+j} & \cdots & \Delta^2 S_i \\ \vdots & & \vdots \\ \Delta^2 S_{i+2j-1} & \cdots & \Delta^2 S_{i+j-1} \end{vmatrix}}$$

with $S_i = 0$ *for* $i < 0$.

Of course we restrict ourselves to the case that the $\varepsilon_j^{(i)}$ are finite; since the ε-algorithm is a nonlinear algorithm, it can always happen that $\varepsilon_{j+1}^{(i)}$ does not exist (when $\varepsilon_j^{(i+1)} - \varepsilon_j^{(i)}$ is singular in Y). It is easy to see now that for $S_i = F_i(x)$, i.e., the partial sums of $F(x) = \sum_{k=0}^{\infty} C_k x^k$, we get

$$\varepsilon_{2m}^{(n-m)} = \frac{\begin{vmatrix} F_n(x) & \cdots & F_{n-m}(x) \\ C_{n+1}x^{n+1} & \cdots & C_{n-m+1}x^{n-m+1} \\ \vdots & & \vdots \\ C_{n+m}x^{n+m} & \cdots & C_n x^n \end{vmatrix}}{\begin{vmatrix} I & \cdots & I \\ C_{n+1}x^{n+1} & \cdots & C_{n+1-m}x^{n+1-m} \\ \vdots & & \vdots \\ C_{n+m}x^{n+m} & \cdots & C_n x^n \end{vmatrix}},$$

The numerator and denominator of $\varepsilon_{2m}^{(n-m)}$ are the determinental formulas for $P(x)$ and $Q(x)$, the solution of the Padé-approximation problem of order (n,m). Let us illustrate this by calculating part of the ε-table for the following nonlinear operator:

$$F\colon C'([1,T]) \to C([1,T])$$

$$\colon x(t) \to e^{x(t)}\frac{dx}{dt} - (1+a),$$

with a a small nonnegative number. The Taylor series expansion is

$$F(x) = \frac{dx}{dt}\sum_{k=0}^{\infty}\frac{1}{k!}[x(t)]^k - (1+a).$$

For the ε-table, we get

$$
\begin{array}{ccccc}
 & 0 & & 0 & \cdots \\
0 & & \dfrac{-1}{1+a} & & \\
 & -(1+a) & & \dfrac{-(1+a)^2}{1+a+\dfrac{dx}{dt}} & \cdots \\
0 & & \dfrac{1}{\dfrac{dx}{dt}} & & \\
 & \dfrac{dx}{dt}-(1+a) & & \dfrac{\dfrac{dx}{dt}-(1+a)(1-x(t))}{1-x(t)} & \cdots \\
0 & & \dfrac{1}{x(t)\dfrac{dx}{dt}} & \vdots & \\
 & \dfrac{dx}{dt}(1+x(t))-(1+a) & \vdots & & \\
 & \vdots & & & \\
0 & & & & \\
\vdots & & & &
\end{array}
$$

3. Applications in operator theory. Several types of nonlinear operator equations

$$F(x)=0$$

can be solved by means of Padé-approximants in operator theory; we mention for instance systems of nonlinear equations, initial value problems, boundary value problems, partial differential equations and nonlinear integral equations. The well-known Newton and Chebyshev iteration [7, pp.205] result respectively from the use of the solution of the Padé-approximation problem of order (1,0) and (2,0) [5], [6]. An interesting new iterative procedure of third order,

$$x_{i+1}=x_i+\frac{\left(-F_i'^{-1}F_i\right)\cdot\left(-F_i'^{-1}F_i\right)}{-F_i'^{-1}F_i+\frac{1}{2}F_i'^{-1}F_i''\left(-F_i'^{-1}F_i\right)^2}$$

where

$F_i=F(x_i)$,

$F_i'=F'(x_i)$ a linear operator (1st Fréchet-derivative at x_i),

$F_i''=F''(x_i)$ a bilinear operator (2nd Fréchet-derivative at x_i),

the division is a multiplication by the inverse element of the denominator,

which we called the Halley iteration [5], [6], proves to be especially interesting in the neighbourhood of singularities because it is derived from the solution of the Padé-approximation problem of order (1, 1). If we use the ε-algorithm for the calculation of the

next iteration step in Halley's method, we have

$$\varepsilon_0^{(0)} = x_i,$$

$$\varepsilon_0^{(1)} = x_i - F_i'^{-1}F_i,$$

$$\varepsilon_0^{(2)} = x_i - F_i'^{-1}F_i - \frac{1}{2}F_i'^{-1}F_i''\left(-F_i'^{-1}F_i\right)^2,$$

$$\varepsilon_1^{(0)} = \left[-F_i'^{-1}F_i\right]^{-1},$$

$$\varepsilon_1^{(1)} = -2\left[F_i'^{-1}F_i''\left(-F_i'^{-1}F_i\right)^2\right]^{-1},$$

$$\varepsilon_2^{(0)} = x_{i+1}.$$

For numerical examples and results we refer to [5], [6].

REFERENCES

[1] C. BREZINSKI, *Algorithmes d'accélération de la convergence*, Editions Technip, Paris, 1978.

[2] _____, *Accélération de la convergence en analyse numérique*, Lecture Notes in Mathematics 584, Springer, Berlin, 1977.

[3] ANNIE CUYT, *Abstract Padé Approximants in Operator Theory*, Lecture Notes in Mathematics 765, Springer, Berlin, 1979, pp. 61–87.

[4] _____, *Regularity and normality of abstract Padé-approximants*, J. Approx. Theory, 35 (1982), pp. 1–11.

[5] ANNIE CUYT AND P. VAN DER CRUYSSEN, *Abstract Padé-approximants for the solution of a system of nonlinear equations*, Comput. Math. Appl., to appear.

[6] ANNIE CUYT, *Padé-approximants in operator theory for the solution of nonlinear differential and integral equations*, Comput. Math. Appl., to appear.

[7] L. RALL, *Computational Solution of Nonlinear Operator Equations*, John Wiley, New York, 1969, reprinted by Krieger, Huntington, New York, 1979.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0016 $01.25/0

A NOTE ON UNIFORM ASYMPTOTIC EXPANSION OF INCOMPLETE LAPLACE INTEGRALS*

K. SONI†

Abstract. An asymptotic expansion of the integral $F(z,a)=\int_a^\infty e^{-z(t-a)}t^{\lambda-1}g(t)\,dt$, $z\to\infty$, which is valid uniformly in $a\geq 0$, is obtained. The technique is elementary, and the expansion is simpler compared to those given earlier by Erdélyi [SIAM J. Math. Anal., 5 (1974), pp. 159–171] and Temme [SIAM J. Math. Anal., 7 (1976), pp. 767–770].

1. Introduction.

Let

$$F(z,a)=\int_a^\infty e^{-z(t-a)}t^{\lambda-1}g(t)\,dt, \tag{1.1}$$

where $a\geq 0$, $0<\lambda<1$, and for some $b\geq 0$, $e^{-bt}g^{(k)}(t)$, $k=0,1,\cdots,n$, are bounded in $[0,\infty)$. In 1974, Erdélyi [2] gave an elegant technique to obtain an asymptotic expansion of $F(z,a)$ which is valid uniformly in $a\geq 0$ as $z\to\infty$. He replaced $t^{\lambda-1}g(t)$ by the fractional integral of a function f defined by

$$f(t)=\begin{cases} g(t), & \lambda=1,\\ (\Gamma(1-\lambda))^{-1}\int_0^t (t-s)^{-\lambda}s^{\lambda-1}g(s)\,ds, & 0<\lambda<1.\end{cases} \tag{1.2}$$

Then, using an integration by parts technique similar to that of Bleistein [1], he obtained the following expansion:

$$F(z,a)=Q\sum_{k=0}^{n-1} z^{-k}\Gamma(k+\lambda)g^{(k)}(0)/k! + \sum_{k=1}^{n-1} z^{-k}I^\lambda f^{(k)}(a)+\mathcal{R}_n, \tag{1.3}$$

where

$$Q=(\Gamma(\lambda))^{-1}\int_a^\infty e^{-z(t-a)}t^{\lambda-1}\,dt, \tag{1.4}$$

$$I^\lambda f(t)=(\Gamma(\lambda))^{-1}\int_0^t (t-s)^{\lambda-1}f(s)\,ds \tag{1.5}$$

and

$$\mathcal{R}_n=z^{1-n}\int_a^\infty e^{-z(t-a)}I^\lambda f^{(n)}(t)\,dt. \tag{1.6}$$

The expansion (1.3) is particularly interesting when the critical point a, which is an end point of the interval of integration, is close to zero, which is a singularity of the integrand. However, in 1976, Temme [3] indicated that the expansion coefficients $I^\lambda f^{(k)}(a)$ are harder to compute and therefore, from a numerical point of view, this expansion is not very attractive. He gave an alternative expansion obtained by expanding $g(t)$ into a power series at $t=a$. The remainder in his expansion has a particularly

*Received by the editors July 23, 1981, and in revised form September 10, 1982.
†Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916.

simple form. The expansion coefficients satisfy a simple recurrence relation, but they are given explicitly as confluent hypergeometric functions with argument az. Therefore, the question whether $F(z,a)$ has an expansion which is uniform in a as $z\to\infty$ and has simpler expansion coefficients, still remains. Erdélyi [2] indicated that Bleistein's procedure of integration by parts can also be used to obtain such an expansion, but he remarked that it does not appear easy to write down explicit expressions for successive terms, to estimate the remainder term or determine the conditions of validity of the resulting expansion. The object of this note is to show how integration by parts can be repeated to obtain an expansion which meets all these requirements. This technique has the advantage that the coefficients as well as the remainder appear in a simpler form, even though, as we will show in the next section, the expansion is in fact the same as that given by Erdélyi. Furthermore, this technique is applicable even when the integrand in (1.1) has logarithmic singularities of positive integral order or when the interval of integration in (1.1) is $(0,a)$.

2. Main result. We assume that a,λ,z are real, $a\geq 0$ and $0<\lambda<1$. $F(z,a)$ is defined by (1.1) where $g\in C^n[0,\infty)$ and $g^{(k)}(t)$, $k=0,1,\cdots,n$, are exponentially bounded (these conditions are the same as in [2]). The functions ϕ and R_k, $k=1,2,\cdots,n$, are defined as follows:

$$\phi(t)=t^{\lambda-1}g(t), \tag{2.1}$$

$$R_k(t)=\phi(t)-\sum_{m=0}^{k-1} c_m t^{m+\lambda-1}, \tag{2.2}$$

where

$$c_m=g^{(m)}(0)/m!. \tag{2.3}$$

This notation is consistent with some recent investigations in asymptotics (see, for example, [4] and [5]). Under the assumptions stated above, $\phi(t)$ and $R_k^{(k)}(t)$, $k=0,1,\cdots,n$, belong to the class $C^n(0,\infty)$. Furthermore, $R_k^{(m)}(t)$, $0\leq m$, $k\leq n$, are exponentially bounded in $[c,\infty)$, $c>0$, and $R_k^{(m)}=O(t^{k-m+\lambda-1})$ as $t\to 0+$.

We prove the following:

THEOREM. *For all z sufficiently large,*

$$F(z,a)=\left(\int_a^\infty e^{-z(t-a)}t^{\lambda-1}\,dt\right)\sum_{k=0}^{n-1} z^{-k}c_k\Gamma(k+\lambda)/\Gamma(\lambda) + \sum_{k=1}^{n} z^{-k}R_k^{(k-1)}(a)+E_n, \tag{2.4}$$

where

$$E_n=z^{-n}\int_a^\infty e^{-z(t-a)}R_n^{(n)}(t)\,dt. \tag{2.5}$$

Proof. By (2.2), $\phi(t)=c_0t^{\lambda-1}+R_1(t)$. Therefore, using integration by parts we obtain

$$\int_a^\infty e^{-zt}\phi(t)\,dt=c_0\int_a^\infty e^{-zt}t^{\lambda-1}\,dt+z^{-1}e^{-az}R_1(a) + z^{-1}\int_a^\infty e^{-zt}R_1'(t)\,dt. \tag{2.6}$$

But by (2.2), for $0<k<n$,

$$R_k^{(k)}(t)=R_{k+1}^{(k)}(t)+c_k(t^{k+\lambda-1})^{(k)}, \tag{2.7}$$

and by using integration by parts

$$\int_a^\infty e^{-zt}R_{k+1}^{(k)}(t)\,dt=z^{-1}e^{-az}R_{k+1}^{(k)}(a)+z^{-1}\int_a^\infty e^{-zt}R_{k+1}^{(k+1)}(t)\,dt. \tag{2.8}$$

Therefore, for $0<k<n$,

$$\begin{aligned}\int_a^\infty e^{-zt}R_k^{(k)}(t)\,dt=&\left(c_k\Gamma(\lambda+k)/\Gamma(\lambda)\right)\int_a^\infty e^{-zt}t^{\lambda-1}\,dt\\&+z^{-1}e^{-az}R_{k+1}^{(k)}(a)+z^{-1}\int_a^\infty e^{-zt}R_{k+1}^{(k+1)}(t)\,dt.\end{aligned} \tag{2.9}$$

The conclusion follows from (2.6) and (2.9). □

Now we will show that the asymptotic expansion obtained above is the same as that given by Erdélyi. The first sum on the right-hand side in (2.4) is identical to the corresponding sum in (1.3). To compare the second sums, we note that

$$\begin{aligned}I^\lambda f^{(k)}(a)=\sum_{m=1}^{k}\frac{a^{\lambda-m}}{(k-m)!}\Bigg[&\frac{(-1)^{m-1}\Gamma(k)\Gamma(m-\lambda)}{\Gamma(m)\Gamma(1-\lambda)}g^{(k-m)}(a)\\&-\frac{\Gamma(k+\lambda-m)}{\Gamma(\lambda-m+1)}g^{(k-m)}(0)\Bigg].\end{aligned} \tag{2.10}$$

This explicit form of $I^\lambda f^{(k)}(a)$ is given by Erdélyi (see [2, §4]). By (2.2) and (2.3),

$$\begin{aligned}R_k^{(k-1)}(t)=&\sum_{m=0}^{k-1}\binom{k-1}{m}\frac{(-1)^m\Gamma(m+1-\lambda)}{\Gamma(1-\lambda)}t^{\lambda-1-m}g^{(k-1-m)}(t)\\&-\sum_{m=0}^{k-1}\frac{\Gamma(\lambda+m)}{\Gamma(\lambda+m-k+1)m!}g^{(m)}(0)t^{\lambda+m-k}.\end{aligned} \tag{2.11}$$

By changing the summation variables in (2.11), we see that $R_k^{(k-1)}(a)=I^\lambda f^{(k)}(a)$. Again, from (1.6) and (2.5), it is not apparent that $\mathfrak{R}_n=E_n$. Nevertheless, the equality holds as a consequence of the corresponding expansions (1.3) and (2.4).

Remark. By Taylor's theorem,

$$R_k(t)=t^{\lambda-1}\left(g(t)-\sum_{m=0}^{k-1}c_mt^m\right)=t^{\lambda-1}(\Gamma(k))^{-1}\int_0^t(t-u)^{k-1}g^{(k)}(u)\,du. \tag{2.12}$$

By successive differentiation, it follows that $I^\lambda f^{(k)}(t)=R_k^{(k-1)}(t)=O(t^\lambda)$ as $t\to 0+$. Since the expression on the right in (2.10) involves negative exponents of a, it is not suitable for the numerical computation of $I^\lambda f^{(k)}(a)$ when a is close to zero. On the other hand, when $g(t)$ is analytic at the origin, we can compute $R_k^{(k-1)}(t)$ from the power series expansion

$$g(t)=\sum_{m=0}^{\infty}c_mt^m.$$

Thus

$$R_k^{k-1}(t)=t^{\lambda}\sum_{m=k}^{\infty}\left(\frac{\Gamma(m+\lambda)}{\Gamma(m+\lambda-k+1)}\right)c_m t^{m-k}. \tag{2.13}$$

If $g(t)$ is not analytic at the origin but has only a finite number of derivatives, say n, in some interval $[0,c]$, we can use (2.12) with some modification. For $0<k\leq n$,

$$R_k(t)=\sum_{m=k}^{n-1}c_m t^{m+\lambda-1}+[\Gamma(n)]^{-1}t^{\lambda-1}\int_0^t(t-u)^{n-1}g^{(n)}(u)\,du. \tag{2.14}$$

We differentiate $R_k(t)$ above $(k-1)$ times and then make a change of variable in the resulting k integrals. Thus,

$$R_k^{(k-1)}(t)=\sum_{m=k}^{n-1}c_m\left[\frac{\Gamma(m+\lambda)}{\Gamma(m+\lambda-k+1)}\right]t^{m+\lambda-k} + t^{n+\lambda-k}\sum_{m=0}^{k-1}K_m\int_0^1(1-u)^{n+m-k}g^{(n)}(tu)\,du, \tag{2.15}$$

where

$$K_m=\binom{k-1}{m}\frac{\Gamma(\lambda)}{\Gamma(\lambda-m)\Gamma(n+m-k+1)}. \tag{2.16}$$

REFERENCES

[1] N. BLEISTEIN, *Uniform asymptotic expansions of integrals with stationary point and nearby algebraic singularity*, Comm. Pure Appl. Math., 19 (1966), pp. 353–370.

[2] A. ERDÉLYI, *Asymptotic evaluation of integrals involving a fractional derivative*, this Journal, 5, (1974), pp. 159–171.

[3] N. M. TEMME, *Remarks on a paper of A. Erdélyi*, this Journal, 7 (1976), pp. 767–770.

[4] K. SONI, *On uniform asymptotic expansions of finite Laplace and Fourier integrals*, Proc. Roy. Soc. Edinburgh, 85A (1980), pp. 299–305.

[5] K. SONI AND R. P. SONI, *A note on uniform asymptotic expansions of finite K_ν and related transforms with explicit remainder*, J. Math. Anal. Appl.,79 (1981), pp. 163–177.

SIAM J. MATH. ANAL.
Vol. 14, No. 5, September 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1405-0017 $01.25/0

THE INVERSE PROBLEM FOR THE VOCAL TRACT AND THE MOMENT PROBLEM*

H. J. LANDAU†

Abstract. B. Gopinath and M. M. Sondhi have shown how the shape of a vocal tract can be determined from an acoustical measurement at the lips. The applications of this method include the synthesis of non-uniform transmission lines and the study of the behavior of the basilar membrane. We show here that the method can be very simply understood as representing an orthogonal decomposition which enters naturally into the classical moment problem.

1. Introduction. In their remarkable solution to the inverse problem for the vocal tract, B. Gopinath and M. M. Sondhi showed that the shape of the tract can be determined from an acoustical measurement at the lips by means of a certain convolution equation [12]. Specifically, assume that wave propagation in the tract is planar and therefore depends only on the cross-sectional areas, expressed as a function $A(x)$ of the distance x from the lips. Let units be chosen so that sound velocity and air density equal 1, and let $V(x) \equiv \int_0^x A(u)du$ be the volume of the tract from the lips to x. Now suppose that a unit impulse of volume velocity is applied at the lips to an initially quiescent tract, and the resulting pressure at the lips is observed as a function of time. This pressure, termed the *impulse response*, has the form $\delta(t) + H(t)$. Intuitively, $H(t)$ for $t \leqslant 2a$ contains information about the tract up to $x = a$, since $2a$ is the time required for the initial impulse to reach the point $x = a$ of the tract, and for a reflection to return. Gopinath and Sondhi determined $A(x)$, $x \leqslant a$, from $H(t)$, $t \leqslant 2a$, in the following way. Assuming $A(x)$ to be positive and continuously differentiable, $0 \leqslant x \leqslant a$, they proved first that $H(t)$ (real) is continuous and that the symmetrized kernel $\delta(t-s) + \frac{1}{2}H(|t-s|)$ is positive definite for $0 \leqslant |s|, |t| \leqslant a$, i.e., that the quadratic form

$$\int_{-a}^{a} g^2(s)ds + \tfrac{1}{2}\int_{-a}^{a} dt \int_{-a}^{a} ds\, g(s)g(t)H(|t-s|) \tag{1.0}$$

is positive for each choice of $g(s)$. This is intuitively plausible because (1.0) is the same as the quadratic form generated by the impulse response, which in turn can be interpreted as the energy stored in the tract. Pursuing positive definiteness a little further, they then modified the kernel to $\delta(t-s) + \frac{1}{2}\{H(|t-s|) - \rho\}$, with some positive number ρ, and considered it on the subinterval $0 \leqslant |s|, |t| \leqslant r \leqslant a$. One can expect this to remain definite for ρ sufficiently small (by continuity) but not for all ρ, the critical value of ρ depending on r. Accordingly, they let

$$\rho_r = \sup\{\rho \mid \delta(t-s) + \tfrac{1}{2}[H(|t-s|)-\rho] \text{ is positive definite for } 0 \leqslant |s|, |t| \leqslant r\}.$$

Finally, they let $f(r,t)$ be the (unique) solution of

$$f(r,t) + \tfrac{1}{2}\int_{-r}^{r} H(|t-s|)f(r,s)ds = 1, \quad |t| \leqslant r \leqslant a. \tag{1.1}$$

Then they proved that the tract cross-section is given uniquely by any of the following relationships, $0 \leqslant r \leqslant a$:

*Received by the editors December 30, 1981. This paper was typeset by Ann Marie McDonough at Bell Laboratories, Murray Hill, New Jersey, using the **troff** program running under the Unix™ operating system.

†Bell Laboratories, Murray Hill, New Jersey 07974.

$$A(r) = f^2(r,r) , \tag{1.2}$$

$$V(r) = \int_0^r f(r,s)ds , \tag{1.3}$$

$$\frac{1}{V(r)} = \rho_r . \tag{1.4}$$

Owing to the analogy between pressure and voltage, this method also solves the problem of synthesizing nonuniform transmission lines [6]. Finally, M. M. Sondhi has recently used a similar approach to probe the behavior of the basilar membrane [13].

These considerations are closely related to work of M. G. Krein who, in a brief note full of striking and general results [8], showed, assuming only that (1.1) with a continuous H has a unique solution for each $0 \leqslant r < a$, that $f(r,t)$ can be transformed to yield the solutions $y(r,\lambda)$ of the differential equation

$$\frac{d}{dr} A(r) \frac{dy}{dr} + \lambda^2 A(r) y(r,\lambda) = 0, \quad 0 \leqslant r < a , \tag{1.5}$$

where $A(r) = f^2(r,r)$. The connection with the vocal tract stems from the fact that, when $A(r)$ is the cross-sectional area, (1.5) is Webster's horn equation, satisfied by a pressure wave of the form $p(r,t) = y(r,\lambda)e^{i\lambda t}$ in the tract.

The conclusions of [6], [12] were derived from the partial differential equations which govern the behavior of pressure and volume velocity in the tract; those of [8], by direct verification. The proofs in both cases are a tour de force of analysis, but offer relatively little to guide the intuition. Specifically, although (1.3) is explained physically as representing conservation of mass, the derivation of (1.2), (1.4) and (1.5) leaves their interpretation and interconnections in mystery. Given the broad usefulness of these results, it seems worthwhile to try to explain them from as many points of view as possible. Here we will show, under positive definiteness assumptions, how that can be done in terms of elementary Hilbert space geometry. Our discussion will be based on one we have found useful in treating the classical moment problem [11]. We should emphasize that the existence of a close analogy between inverse problems for Sturm-Liouville equations and the moment problem was repeatedly pointed out by M. G. Krein [9], [10], albeit without details. Our contribution therefore consists of making this connection explicit, and stressing a geometric interpretation. This makes the mathematical relationships (1.1)—(1.5) easy to understand, and shows that they depend only on the positivity of (1.0), not on the partial differential equations which connect it to a specific physical problem. These equations are necessary, however, for the illuminating identification in [6], [12] of $\delta(t) + H(t)$ with the impulse response, which opens the path to applications.

The classical moment problem consists of asking whether a prescribed sequence of numbers $1 = s_0, s_1, \cdots$ can be represented in the form

$$s_k = \int_{-\infty}^{\infty} x^k d\mu(x), \qquad k \geqslant 0,$$

with some positive measure $d\mu(x)$, generally supposed to have an infinite number of points of increase. Clearly a necessary condition is that the quadratic form

$$\sum\sum q_j \bar{a}_k s_{j+k} = \int_{-\infty}^{\infty} |\sum a_k x^k|^2 \, d\mu(x)$$

be positive for every choice of a finite number of $\{a_j\}$; this turns out also to be sufficient. The trigonometric moment problem asks for representation in the form

$$s_k = \int_{\theta}^{2\pi} e^{ik\theta}\, d\mu(\theta), \qquad d\mu \geqslant 0,\ k \geqslant 0,$$

and has a corresponding solution. Together with their variants, these representations have illuminated an extraordinary range of subjects: a partial list includes analytic and harmonic functions, the spectral theory of operators, prediction theory, approximation, numerical analysis. In a very vague sense, some notion of positivity is the unifying feature underlying these applications.

Somewhat more specifically, there is in many situations a positive definite quadratic form defined on an increasing family of subspaces. Thus in the moment problem the quadratic form is

$$\sum s_{j+k}\, a_j \bar{a}_k \tag{1.6}$$

and the subspaces can be thought of as the sequences $\{a_k\}$, $k \leqslant N$, of given length N, with the length successively increasing. Similarly, in a mechanical or electrical system, the positive form might be energy, and the subspaces composed of functions of time, representing possible forces applied to the system, and of duration T, with T increasing. Now a familiar positive quadratic form is the scalar product in a Hilbert space H. Moreover, if we can further identify elements v of H with functions

$$v \leftrightarrow f_v(x) \tag{1.7}$$

in such a way that the scalar product of v and $w \in H$ is given by

$$\int f_v(x) f_w(x)\, d\mu(x)\ , \tag{1.8}$$

for some positive measure $d\mu$, we will have a particularly clear view of the behavior of the quadratic form, since it is now exhibited as acting independently on the components of $f_v(x)$ obtained by restricting $f_v(x)$ to disjoint subsets of its domain of definition. It is the general objective of spectral decomposition to produce representations of this kind. We remark that even when the elements v of H are themselves functions, the associated f_v does not in general coincide with v.

We can now observe that the moment problem poses this very question: it asks that the quadratic form (1.6) be expressed by (1.8), with the correspondence (1.7) given by

$$\{a_0,\cdots,a_N\} \leftrightarrow \sum_{k=0}^{N} a_k x^k\ .$$

In [11], we employed the positivity of the quadratic form given by $s_0,\cdots,s_{2n}$ to define a scalar product for polynomials in x of degree n. In this space of polynomials, we selected a basis, consisting of polynomials of successively increasing degree; these satisfy a three-term recursion which can be viewed as a discrete analogue of a second-order Sturm-Liouville differential operator. We then considered the linear operation which assigns to each polynomial its value at a given point. Much of the fundamental information about the moment problem flowed easily from properties of this operation and its interplay with the basis.

The same considerations apply in the present instance. In order to avoid certain purely technical complications of the continuous case, we will first consider the

problem in a discrete form, so as to pose it in a finite-dimensional space, where it will become transparent; in this respect, we follow the approach of recent papers, of K. M. Case and M. Kac [4], [5]. The discrete analogue has also been studied in [3], and a detailed treatment of it, focusing on efficient computation, is given in [1]. Finally, the general point of view we adopt has been discussed in the rich paper [7]. In outline,

1. We discretize (1.1) in such a way that the Toeplitz matrix C which replaces the operator on the left-hand side is positive definite, and use this matrix to define a scalar product $[\cdot,\cdot]$ on spaces Π_N of even trigonometric expressions of the form $S_N(z) = \sum_{-N}^{N} a_m z^m$, $z = e^{i\delta\lambda}$.
2. We select a natural orthonormal basis for Π_n, consisting of trigonometric expressions of successively increasing degree.
3. We focus on the element $E_N \in \Pi_N$ (sometimes called a *reproducing kernel*), defined by $[S_n,E_n] = S_N(1)$ for each $S_N \in \Pi_N$.
4. We show that the discrete analogues of (1.2)–(1.5) are immediate consequences of expressing E_N in terms of the basis.
5. We obtain the original continuous versions either by passing to the limit as the discretization becomes finer, or by applying our reasoning in its continuous form from the outset.
6. Finally, we show that the correspondence between $H(t)$ and $A(x)$ described here is unique.

We proceed to take up these topics in turn, commenting more extensively on motivation as appropriate.

2. A scalar product. Let a,b denote $(K+1)$-dimensional vectors with components $\{a_m\}$, $\{b_m\}$, respectively, and (a,b) the usual scalar product,

$$(a,b) = \sum_{m=o}^{K} a_m \bar{b}_m .$$

Let C be a $(K+1) \times (K+1)$ positive definite matrix, i.e., one for which $(a,Ca) > 0$ except when $a = 0$. In view of the ultimate application to (1.1), we will assume C to be real. Then C is symmetric and, by virtue of its positive definiteness, the quadratic form

$$[a,b] \equiv (a,Cb) \tag{2.1}$$

can be viewed as defining another scalar product on the vectors. Suppose C is also a Toeplitz matrix, i.e., one whose entry $c_{m,n}$ depends only on $(m-n)$. Then it is convenient to associate with a vector a the trigonometric polynomial

$$S_a(e^{i\delta\lambda}) = \sum_{m=0}^{K} a_m e^{im\delta\lambda}$$

with some $\delta > 0$, and to think of the scalar product as defined on the corresponding polynomials,

$$[S_a,S_b] \equiv [a,b] , \tag{2.2}$$

for in this way the Toeplitz property is succinctly expressed by

$$c_{m-n} = [e^{im\delta\lambda}, e^{in\delta\lambda}] . \tag{2.3}$$

The form of this expression in turn suggests the trigonometric moment problem, i.e.,

the problem of finding a positive measure $d\mu(\lambda)$ for which

$$c_m = \int_{-\pi/\delta}^{\pi/\delta} e^{im\delta\lambda} d\mu(\lambda) , \quad 0 \leqslant m \leqslant 2K . \tag{2.4}$$

The scalar product $[S_a, S_b]$ on trigonometric polynomials can serve as the point of departure for a discussion of this question, and we will return to some aspects of it later. For the time being, however, we take a slightly different course, and use C to define a scalar product for two-sided trigonometric expressions.

Specifically, if $K = 2N$ then, with $z = e^{i\delta\lambda}$, C will define, by means of (2.3), a scalar product in the space of functions of the form $\sum_{-N}^{N} a_m z^m$, $|\lambda| \leqslant \pi/\delta$. Since C is symmetric, even and odd functions of λ are orthogonal with respect to $[\cdot,\cdot]$, and we concentrate on the former, denoting by Π_N the $(N+1)$-dimensional space of functions of the form $\sum_{-N}^{N} a_m z^m$, $a_{-m} = a_m$, $|\lambda| \leqslant \pi/\delta$. Similarly, if C has even order $(K = 2N + 1)$, on setting

$$c_{m-n} = \left[e^{i(m+1/2)\delta\lambda} , e^{i(n+1/2)\delta\lambda} \right] ,$$

we can define a scalar product in the space Π_N^e of even functions of the form $\sum_{m=0}^{N} a_m \, (e^{i(m+\frac{1}{2})\delta\lambda} + e^{-i(m+\frac{1}{2})\delta\lambda})$. We henceforth focus on Π_N. As we shall see, despite the fact that C is a Toeplitz matrix, this formulation will lead us to considerations which more closely resemble the power moment problem, generated by Hankel matrixes. This stems from the fact that members of Π_N have the form $\sum_{0}^{N} 2a_m \cos m\delta\lambda$, hence are polynomials of degree N in $\cos \delta\lambda$.

As a matter of notation, if $T_K(z) = \sum_{-K}^{K} \tau_m z^m$ is an element of Π_K, we will refer to τ_K as the *leading coefficient* of T_K, and to the vector $(\tau_{-K},\cdots,\tau_K)$ as the *coefficient vector* of T_K. Thus if T_a and T_b are elements of Π_N with coefficient vectors a and b respectively, we have as in (2.1) and (2.2)

$$[T_a, T_b] = (a, Cb) , \tag{2.5}$$

or explicitly

$$[T_a, T_b] = \sum_{j,m=-N}^{N} a_j \bar{b}_m c_{j-m} . \tag{2.6}$$

Let us observe that, although the scalar product is not necessarily defined in Π_{N+1}, nevertheless by (2.6)

$$[zS, zT] = [S, T]$$

so that

$$[(z+z^{-1})S, T] = [S, (z+z^{-1})T], \tag{2.7}$$

whenever either scalar product is defined, in particular if S or $T \in \Pi_{N-1}$. Finally, we set

$$\|T\|^2 = [T, T] .$$

3. A basis. The matrix C, being symmetric and Toeplitz, is generated by the $2N+1$ entries $c_0, c_1, \cdots, c_{2N}$. If only the first $2K+1$ of these entries are given, the corresponding matrix defines the scalar product on Π_K, and as additional c_i are added in pairs, this scalar product is extended to larger subspaces $\Pi_{K+1} \subset \cdots \subset \Pi_N$, without being altered where previously specified. Thus it is natural, in choosing a basis for

Π_N, to select the basis vectors $\{P_m\}$ so as to span, successively, the chain of subspaces $\Pi_0 \subset \Pi_1 \subset \cdots \subset \Pi_{N-1} \subset \Pi_N$. This can be done by the Gram-Schmidt process. Specifically, let C_K denote the principal $(2K+1) \times (2K+1)$ minor of C, and suppose that $Q_K(z) \in \Pi_K$ has coefficient vector q with leading coefficient 1 and such that

$$C_K\, q = \left(\frac{\nu_K^2}{2}, 0, \cdots, 0, \frac{\nu_K^2}{2} \right). \tag{3.1}$$

Then by (2.5), $[T_{K-1}, Q_K] = 0$ for each $T_{K-1} \in \Pi_{K-1}$, hence Q_K is a scalar multiple of P_K. Again by (2.5), $\|Q_K\|^2 = (q, Cq) = \nu_K^2$, and so

$$P_K(z) = \frac{Q_K(z)}{\nu_K}. \tag{3.2}$$

By Cramer's rule applied to (3.1), the coefficients of P_K are all real.

4. Evaluations. Looking briefly back to the moment problem, if $d\mu$ is a measure satisfying (2.4), then from (2.6) and (2.3)

$$\|T_a\|^2 = \int_{-\pi/\delta}^{\pi/\delta} |T_a(e^{i\delta\lambda})|^2 \, d\mu(\lambda),$$

so that this problem asks, in effect, how the norm of an element in Π_N is related to its values. Viewing the question in this form, it is natural to start with evaluation at $\lambda = 0$, i.e., at $z = 1$. Since the map of $T_K \in \Pi_K$ into $T_K(1)$ is a linear functional, there exists a unique *evaluation element* $E_K(z) \in \Pi_K$ (sometimes called a *reproducing kernel*) such that

$$[T_K, E_K] = T_K(1), \tag{4.1}$$

for every $T_K \in \Pi_K$. The next proposition is immediate.

PROPOSITION 1. *The following characterizations of $E_K(z)$ are equivalent:*

a) $$E_K(z) = \sum_{m=0}^{K} P_m(1) P_m(z); \tag{4.2}$$

b) *the coefficient vector e of E_K satisfies*

$$Ce = (1,1,\cdots,1); \tag{4.3}$$

c) *$E_K(z)$ is the solution to the extremal problem*

$$\max_{T \in \Pi_K} \frac{|T(1)|}{\|T\|}. \tag{4.4}$$

Proof. Since the coefficients of P_m are real, $\overline{P_m(1)} = P_m(1)$. When E is defined by (4.2) we have, by orthonormality of $\{P_j\}$, $[P_m, E_K] = P_m(1)$ for each $m \leqslant K$, hence also for all linear combinations of these $\{P_m\}$. Since $T(1)$ coincides with the sum of the coefficients of T, condition (b) follows from (2.5). Finally, by Schwarz's inequality,

$$|T(1)| = |[T, E_K]| \leqslant \|T\| \, \|E_K\|,$$

so that

$$\frac{|T(1)|}{\|T\|} \leqslant \|E_K\| ,$$

with equality only if T is a scalar multiple of E_K. This concludes the proof.

Now by definition, we see that

$$\|E_K\|^2 = [E_K, E_K] = E_K(1) = \sum_{m=0}^{K} P_m^2(1) . \tag{4.5}$$

The quantity $\|E_K\|^{-2}$ is called the *central mass* for C_K because, as is not hard to see, it corresponds to the largest mass that can be concentrated at $\lambda = 0$ by a measure which represents the scalar product as in (2.4) [11]. Equivalently to (4.4), we have the following extremal property for the central mass. Let J_K be the $(2K+1) \times (2K+1)$ matrix with all entries 1, and denote by $C_K(\rho)$ the matrix $C_K - \rho J_K$. Then for $T \in \Pi_K$

$$(C_K(\rho)T,T) = \|T\|^2 - \rho|T(1)|^2 ,$$

so that, by (4.4),

$$\|E_K\|^{-2} = \sup \{\rho | C_k(\rho) \text{ is positive definite}\} . \tag{4.6}$$

Next, by (2.5) and (3.1), the leading coefficient ϵ_K of E_K satisfies

$$\epsilon_K = \left[E_K, \frac{Q_K}{\nu_K^2}\right] ,$$

and so, using the definition of E_K and (3.2) we find

$$\epsilon_K = \frac{1}{\nu_K} P_K(1) . \tag{4.7}$$

Finally, if $T_{K-1} \in \Pi_{K-1}$, then $(z+z^{-1}-2)T_{K-1} \in \Pi_K$ and vanishes at $z = 1$. Thus by (2.7) and the definition of E_K,

$$\left[T_{K-1},(z+z^{-1}-2)E_K\right] = \left[(z+z^{-1}-2)T_{K-1},E_K\right] = 0 . \tag{4.8}$$

Consequently $(z+z^{-1}-2)E_K(z)$, which is an element of Π_{K+1}, is orthogonal to Π_{K-1}, and therefore it must coincide with a linear combination of $P_K(z)$ and $P_{K+1}(z)$. Since it also vanishes at $z = 1$, we have

$$(z+z^{-1}-2)E_K(z) = \alpha_K\{P_{K+1}(z)\ P_K(1) - P_K(z)\ P_{K+1}(1)\} , \tag{4.9}$$

and by using (3.2) and (4.7) to compare the leading coefficients on both sides of (4.9), we find $\alpha_K = \nu_{K+1}/\nu_K$, whence

$$(z+z^{-1}-2)E_K(z) = \frac{\nu_{K+1}}{\nu_K} P_K(1)\ P_{K+1}(1) \left\{\frac{P_{K+1}(z)}{P_{K+1}(1)} - \frac{P_K(z)}{P_K(1)}\right\} . \tag{4.10}$$

This is the Christoffel-Darboux formula, here derived very simply. We rewrite it so as to resemble (1.5) by introducing, for a sequence $\{T_K\}$, the difference operators

$$\Delta_+ T_K = T_{K+1} - T_K ,$$
$$\Delta_- T_K = T_K - T_{K-1} ,$$

and by setting

$$\omega_k = \frac{\nu_{K+1}}{\nu_K} \frac{P_{K+1}(1)}{P_K(1)},$$

$$\psi_K(z) = \frac{\Delta_- E_K}{P_K^2(1)} = \frac{P_K(z)}{P_K(1)}.$$

With these definitions, we see that (4.10) becomes

$$\omega_K P_K^2(1)\Delta_+\psi_K(z) + (2-z-z^{-1})E_K(z) = 0 \tag{4.11}$$

and, on applying Δ_-, we find

$$\Delta_-\omega_K P_K^2(1)\Delta_+\psi_K(z) + (2-z-z^{-1})P_K^2(1)\psi_K(z) = 0. \tag{4.12}$$

We can view this as the familiar three-term recursion satisfied by orthogonal polynomials.

5. The continuous version. We can recognize in (4.3), (4.7), (4.5), (4.6), and (4.12) the analogues of (1.1)-(1.5), respectively. Specifically, writing for brevity

$$(I+H_r)f \equiv f(t) + \tfrac{1}{2}\int_{-r}^{r} h(|t-s|)f(s)ds, \qquad |t| < r,$$

let us set $\delta = 2a/2N+1$, subdivide the interval $|t| \leqslant a$ into $2N+1$ equal subintervals $\{I_m\}, |m| \leqslant N$, and approximate $I+H_a$ by the Toeplitz matrix C having entries

$$c_j = d_j + \tfrac{1}{2}\int_{I_j} h(|u|)du, \qquad |j| \leqslant N,$$

where $d_0 = 1$ and $d_j = 0$ when $j \neq 0$. Then we can show that C is positive definite for all sufficiently large N, with an inverse bounded independently of N, and that the constants ν_K of (3.1) approach $\sqrt{2}$ as $N \to \infty$. If $r < a$ has the form $r = (K+\tfrac{1}{2})\delta$ for some integer K, e is the coefficient vector of $E_K(z)$, and $g_r(t)$ the piecewise constant function, defined for $|t| < r$, having value e_m on I_m, $|m| \leqslant K$, then we can also show that $g_r(t) \to f(r,t)$ uniformly as $N \to \infty$. As, by (4.7), $P_K(1)/\nu_K$ coincides with the leading coefficient $g_r(K\delta)$ of e, it therefore approaches $f(r,r)$ as $N \to \infty$, and likewise $P_m(1)/\nu_m \to f(m\delta,m\delta)$, uniformly for $|m| \leqslant N$. The relation

$$\int_{-r}^{r} f(r,t)dt = 2\int_{0}^{r} f^2(t,t)dt$$

expressed by (1.2) and (1.3) is merely a limiting form of the identity

$$\delta E_K(1) = \delta\|E_K\|^2 = \delta\sum_{m=0}^{K} P_m^2(1).$$

In our discrete approximation, the integral operator defining ρ_r becomes $C_K(\delta\rho/2)$, and we see from (4.6) that this is positive definite if and only if $\delta\rho/2 < \|E_K\|^{-2}$; the limiting form of this is (1.4). Finally, $E_K(z) = \sum_{-K}^{K} e_m z^m = \sum_{-K}^{K} e_m e^{im\delta\lambda}$, so that $\delta E_K(z) \to \int_{-r}^{r} f(r,t)e^{it\lambda}\,dt$, and when f is sufficiently smooth we can therefore expect that

$$\Delta_- E_K(z) \to \frac{d}{dr}\int_{-r}^{r} f(r,t)e^{it\lambda}\,dt \qquad \text{as } N \to \infty .$$

Thus the analogue of $\psi_K(z)$ is

$$\left(\frac{d}{dr}\int_{-r}^{r} f(r,t)e^{it\lambda}dt\right)/2f^2(r,r) ,$$

and since $\omega_K \to 1$ while $(2-z-z^{-1}) = 2-2\cos\delta\lambda \to \delta^2\lambda^2$, (4.12) assumes the limiting form (1.5).

While the approach just described explains the origin of the results and their interrelation, we can give actual proofs more efficiently by a slightly different path, which follows the geometric line of reasoning in its continuous version from the outset. Here instead of the even trigonometric expressions of Π_K we consider even functions of the form

$$G_r(\lambda) = \int_{-r}^{r} g(s)e^{is\lambda}ds , \tag{5.1}$$

with g (the continuous analogue of the coefficient vector) an even function in $L^2(-r,r)$; consequently by the Paley-Wiener theorem, Π_K is replaced by the space Ω_r of even functions of exponential type r, square-integrable on the reals. We will establish the following correspondence which, as we have seen, leads directly to (1.1)-(1.5):

	exponential type r	degree K
	Ω_r	Π_K
	$E(r,\lambda) \equiv \int_{-r}^{r} f(r,t)e^{it\lambda}dt$	evaluation element $E_K(z)$
(5.2)	$V(r) \equiv \tfrac{1}{2}\|E(r,\lambda)\|^2$	
	$A(r) \equiv f^2(r,r)$	$\tfrac{1}{2}P_k^2(1)$
	$dE(r,\lambda)/dr$	$P_K(z)P_k(1)$
	Equation (1.5)	Christoffel-Darboux formula.

Although the continuous case inevitably presents certain technical difficulties, the basic line of argument follows exactly that of the discrete problem and so we relegate it to the appendix.

At this point we see that to each positive quadratic form (1.0) there correspond the quantities and relationships (1.1)-(1.5). We will show next that the correspondence between H and A is one-to-one. All of this is independent of the origin of (1.0). To interpret these quantities physically, we can view (1.5) as Webster's horn equation, associated with a vocal tract of cross-sectional area $A(x)$. Then $V(r)$ becomes the volume of the tract to r, and (1.3) shows, by the conservation argument of [6], [12], that $f(r,t)$ coincides with that volume velocity required at $x=0$ (as a function of time) to produce, at $t = r$, a unit pressure for $0 \leqslant x \leqslant r$ along the tract. Now, as proved in [6], [12], it follows from the partial differential equations which relate volume velocity to pressure that this is the same as the even version of the volume velocity which produces a unit pressure at $x = 0$, $0 \leqslant t \leqslant r$. In turn, (1.1) shows that the kernel $\delta(t) + H(t)$ should be viewed as that which transforms volume velocity to pressure at $x=0$, hence as the impulse response of the tract.

A comprehensive discussion of the connection between a variety of inverse problems and integral equations is contained in [2].

6. Uniqueness. We wish to show that the functions $H(s)$, $0 \leqslant s \leqslant 2a$, and $A(r)$, $0 \leqslant r \leqslant a$, which are related by (1.1), (1.2) and (1.5) determine each other uniquely. We begin with the discrete problem which is of interest in its own right. For that case, uniqueness was proved in [3], by a different method. We will show that our present point of view again suggests a simple approach.

We have seen that the $2N+1$ entries $c_0,\cdots,c_{2N}$ of the matrix C define a scalar product on Π_N and this in turn determines the central mass sequence or equivalently $\{\|E_m\|^2\}$, $m = 0,\cdots,N$. The same matrix likewise generates a scalar product in Π_N^e and a corresponding sequence $\{\|E_m^e\|^2\}$, $m = 1,\cdots,N$. To show that these $2N+1$ quantities taken together in turn determine C, we return to the original point of view, in which by means of (2.3) $c_0,\cdots,c_N$ define a scalar product on the space Π_N^* of trigonometric polynomials $S(z) = \sum_{m=0}^{N} \sigma_m z^m$, $z = e^{i\delta\lambda}$. Proceeding just as before, we introduce the basis of (Szegö) orthogonal polynomials $\{T_K(z)\}$, $0 \leqslant K \leqslant N$,

$$T_K(z) = \frac{U_K(z)}{\mu_K}, \tag{6.1}$$

where the coefficient vector u of U_K has leading coefficient 1, and satisfies

$$C_K^* u = (0,0,\cdots,\mu_K^2), \tag{6.2}$$

with C_K^* the $(K+1) \times (K+1)$ principal minor of C. We then consider $E_K^*(z)$, defined by

$$[S_K, E_K^*] = S_K(1),$$

for each $S_K \in \Pi_K^*$. By the reasoning of Proposition 1, we find

$$\|E_K^*\| = \max_{S_K \in \Pi_K^*} \frac{|S_K(1)|}{\|S_K\|}. \tag{6.3}$$

Now if $K = 2L$, let us consider $V_K(z) \equiv S_K(z) z^{-L}$. This is now a two-sided trigonometric expression and we see that $|S_K(1)| = |V_K(1)|$, $\|S_K\| = \|V_K\|$. Moreover, in the decomposition

$$V_K(z) = V_e(z) + V_o(z),$$

into components V_e and V_o which are even and odd functions of λ, respectively, V_o vanishes at $\lambda = 0$, i.e., $V_K(1) = V_e(1)$, while V_e and V_o are orthogonal in our scalar product, so that $\|V_K\|^2 = \|V_e\|^2 + \|V_o\|^2 \geqslant \|V_e\|^2$. It follows that in (6.3) we can suppose $V_K = V_e$, whereupon $V_K \in \Pi_L$, and we find that $\|E_K^*\| = \|E_L\|$. Analogously, for K odd, $K = 2L+1$, $\|E_K^*\| = \|E_{L+1}^e\|$. We conclude that the known sequences $\{\|E_m\|^2\}$, $0 \leqslant m \leqslant N$ and $\{\|E_m^e\|^2\}$, $1 \leqslant m \leqslant N$ together determine $\{\|E_m^*\|^2\}$, $0 \leqslant m \leqslant 2N$, or, equivalently, the sequence $\{T_K^2(1)\}$, $0 \leqslant K \leqslant 2N$.

PROPOSITION 2. *The sequence* $\{T_m^2(1)\}$, $0 \leqslant m \leqslant K$ *uniquely determines the orthogonal polynomials* $\{T_m(z)\}$, $0 \leqslant m \leqslant K$.

Proof. By definition

$$[z^j, U_K(z)] = 0, \quad 0 \leqslant j \leqslant K-1, \tag{6.4}$$

and so, from (2.3),

$$[z^{j+1}, zU_K(z)] = 0\,, \quad 0 \leqslant j \leqslant K-1\,. \tag{6.5}$$

Now $zU_K(z) \in \Pi_{K+1}^*$ has leading coefficient 1, and the same is true of $U_{K+1}(z)$, which is orthogonal to all of Π_K^*. Thus $zU_K(z) - U_{K+1}(z) \in \Pi_K^*$ and from (6.5)

$$[z^j, zU_K(z) - U_{K+1}(z)] = 0\,, \quad 1 \leqslant j \leqslant K\,. \tag{6.6}$$

On taking complex conjugates in (6.4) and applying (2.3) we discover

$$[z^K z^{-j}, z^K \overline{U_K(z)}] = 0\,, \quad 0 \leqslant j \leqslant K-1$$

so that the polynomial $W_K(z) = z^K \overline{U_K(z)} \in \Pi_K^*$ likewise satisfies the orthogonality relations (6.6). As these define a one-dimensional subspace of Π_K^*, we conclude that

$$zU_K(z) - U_{K+1}(z) = \alpha_K W_K(z)\,, \tag{6.7}$$

for some α_K, which is real since all the coefficients in (6.7) are. This relation is well-known.

We see from the definition of $W_K(z)$ that its coefficient vector is found from that of $U_K(z)$ by writing the components in reverse order. Consequently, $\|W_K\| = \|U_K\|$ and $W_K(1) = U_K(1)$. We can therefore find the norms on both sides of (6.7), rewritten as $zU_K(z) = U_{K+1}(z) + \alpha_K W_K(z)$, and also evaluate (6.7) at $z = 1$; in the former computation, by (2.3) and (6.1), $[zU_K, zU_K] = [U_K, U_K] = \mu_K^2$, and by definition of U_{K+1}, $[W_K, U_{K+1}] = 0$. Thus we obtain

$$\mu_K^2 = \mu_{K+1}^2 + \alpha_K^2\, \mu_K^2\,, \tag{6.8}$$

$$\mu_K T_K(1) - \mu_{K+1} T_{K+1}(1) = \alpha_K \mu_K T_K(1)\,. \tag{6.9}$$

From (6.8),

$$(1 - \alpha_K^2) = \frac{\mu_{K+1}^2}{\mu_K^2}\,, \tag{6.10}$$

while from (6.9)

$$\frac{T_K(1)}{T_{K+1}(1)}\,(1-\alpha_K) = \frac{\mu_{K+1}}{\mu_K}\,. \tag{6.11}$$

Since $\mu_K^2 = \|U_K\|^2 > 0$, $|\alpha_K| < 1$ by (6.10), and combining (6.10) and (6.11) yields

$$\alpha_K = \frac{T_K^2(1) - T_{K+1}^2(1)}{T_K^2(1) + T_{K+1}^2(1)}\,. \tag{6.12}$$

We conclude that the sequence $\{T_m^2(1)\}$, $0 \leqslant m \leqslant K$ determines the quantities $\{\alpha_m\}$, $0 \leqslant m \leqslant K-1$ by (6.12), as well as $\{\mu_m\}$, $0 \leqslant m \leqslant K$ by (6.11). Thus using $\{\alpha_m\}$, we can generate $\{U_m(z)\}$ by (6.7), $0 \leqslant m \leqslant K$, and renormalize by the known $\{\mu_m\}$ to obtain $\{T_m(z)\}$. This completes the proof.

To return to the original problem, given the increasing sequence $\{\|\mathrm{E}_m^*\|^2\}$, we take successive differences to find $\{T_m^2(1)\}$, generate the polynomials $\{U_m(z)\}$ as in Proposition 2, and these in turn successively determine the matrix entries, from the equation for the first component of (6.2). But even more generally, starting with a sequence of positive numbers $\{p_K\}$, which we do not know a priori to correspond to values $\{T_m^2(1)\}$ for some set of orthogonal polynomials, we can check that the above

construction generates a matrix C which is positive definite and for which $p_m = T_m^2(1)$.

In the continuous version, we can expect the distinction between $\|E_m\|$ and $\|E_m^e\|$ to disappear. Indeed, given $A(r)$ positive and continuously differentiable for $0 \leqslant r \leqslant a$, we solve (1.5) as an initial value problem for $y(r, \lambda)$, determine $E(r, \lambda) = \int_0^r A(s)y(s,\lambda)ds$, and thereby also its inverse Fourier transform $f(r,t)$. Then (1.1) evaluated at $t = r$ yields

$$f(r,r) + \tfrac{1}{2} \int_0^{2r} H(u)f(r, r-u)du = 1$$

and on differentiating this with respect to r we obtain, analogously to the discrete case, a Volterra equation which we can solve for $H(u)$, $0 \leqslant u \leqslant 2a$. We can also give another description of H, which draws on spectral theory for differential equations. Let us consider the $\{\lambda_K\}$ for which $y(a, \lambda_K) = 0$. The corresponding $\{\sqrt{A(r)}y(r, \lambda_K)\}$, being eigenfunctions of a self-adjoint boundary value problem, form a complete orthogonal set in $L^2(0,a)$. Let $n_K^2 = \int_0^a A(r)y^2(r,\lambda_K)dr$ and let $d\mu(\lambda)$ be the measure which at $\lambda = \lambda_K$ has the mass $1/n_K^2$. By expanding a function $g(r)$ in the normalized eigenfunctions, we can see that the transformation $G(\lambda) = \int_0^a g(r)\sqrt{A(r)}y(r, \lambda)dr$ maps $L^2(0, r)$ unitarily onto $L^2(d\mu)$.

Specifically, as $\{G(\lambda_K)/n_k\}$ are the coefficients of $g(t)$ in the expansion, we have

$$g(r) = \sum_K G(\lambda_K) \frac{\sqrt{A(r)}\, y(r, \lambda_K)}{n_K^2} = \int G(\lambda)\sqrt{A(r)}y(r, \lambda)d\mu(\lambda) ,$$

and

$$\int_0^a g^2(r)dr = \sum \frac{G^2(\lambda_K)}{n_K^2} = \int G^2(\lambda)d\mu(\lambda) .$$

These relations also show that any element in $L^2(d\mu)$, i.e., any sequence of values square-summable with respect to the weights n_K^{-2}, is representable in the form $\{G(\lambda_K)\}$ for some $g(r)$. Now if we set

$$E(\lambda) = \int_0^a \sqrt{A(r)}\, \sqrt{A(r)}\, y(r,\lambda)dr ,$$

we see that $E(\lambda)$ effects evaluation at $\lambda = 0$ in the scalar product of $L^2(d\mu)$, for by unitarity of the expansion,

$$\int G(\lambda)\overline{E(\lambda)}d\mu(\lambda) = \int_0^a g(r)\, \sqrt{A(r)}\, dr = G(0) .$$

Since the inverse Fourier transform of $E(\lambda)$ is $f(a,t)$, while that of $E(\lambda)d\mu(\lambda)$ is 1 by the evaluation property, we can identify the kernel $\delta(t) + \tfrac{1}{2} H(t)$ with the Fourier transform of the measure $d\mu(\lambda)$. We omit the details.

7. Appendix. We devote this section to establishing the correspondence (5.2).

Since $I+H_a$ is positive definite, we can, analogously to (2.6), define for $F,G \in \Omega_a$

$$(7.1)\quad [G,F] \equiv (g,(I+H_a)f) = \int_{-\infty}^{\infty} g(t)\left(\overline{f(t)} + \tfrac{1}{2}\int_{-\infty}^{\infty} H(|t-u|)\overline{f(u)}\,du\right)dt\ .$$

In these integrals, f and g, being the Fourier transform of F,G, respectively, vanish identically outside $(-a,a)$; thus [G,F] depends on the values of $H(|x|)$ in the interval $|x| \leqslant 2a$. These are insufficient to define the scalar product in Ω_s for $s > a$. Nevertheless by (7.1) they do determine $[G, F]$ when $G \in \Omega_s$ and $F \in \Omega_{2a-s}$, since in this case the values of $|t-u|$ entering into (7.1) do not exceed $2a$. If $G(\lambda) \in \Omega_{a-\epsilon}$ and $Q(\lambda) \in \Omega_\epsilon$, with Fourier transforms $g(t)$ and $q(t)$ respectively, the product $G(\lambda)Q(\lambda)$ corresponds under Fourier transformation to the convolution of $g(t)$ with $q(t)$, so that $G(\lambda)Q(\lambda) \in \Omega_a$, and from (7.1) we see that, as in (2.7),

$$(7.2)\qquad [GQ, F] = [G, \bar{Q}F]\ .$$

Finally, if $f_n \to f$ in $L^2(-a,a)$ or equivalently $F_n(\lambda) \to F(\lambda)$ in $L^2(-\infty,\infty)$, then by (7.1) also

$$(7.3)\qquad [G, F_n] \to [G, F]\ .$$

Next, if $f(r,t)$ satisfies (1.1), and we set

$$(7.4)\qquad E(r,\lambda) = \int_{-r}^{r} f(r,t)e^{it\lambda}dt\ ,$$

we see that $E(r, \lambda) \in \Omega_r$ is the evaluation at $\lambda = 0$, since by (7.1) for every $G \in \Omega_r$,

$$(7.5)\qquad [G, E] = G(0)\ .$$

We note that if the kernel H in (1.1) is continuous, $f(r,t)$ is continuously differentiable in both variables. For on differentiating (1.1) with respect to t we find

$$(I+H_r)f_2(r,t) = f(r, r)[H(r-t)-H(r+t)]\ ,\quad |t| < r\ ,$$

with $f_2(r,t) \equiv \partial f(r,t)/\partial t$. Since $(I+H_r)$ has a bounded inverse, $f_2(r,t)$ is square-integrable, whereupon $H_r f_2(r,t)$ is continuous, so that writing $f_2(r,t) = f(r,r)[H(r-t)-H(r+t)] - H_r f_2$ we see that likewise $f_2(r,t)$ is continuous. A similar argument applies to $f_1(r,t)$. Moreover, $f(r, r) \neq 0$, else $f_2(r, t) = 0$, so that $f(r, t) = const = f(r, r) = 0$," a contradiction.

We can now immediately derive (1.1)-(1.4). For on applying definitions and (7.5),

$$\|E(r, \lambda)\|^2 \equiv [E(r, \lambda), E(r, \lambda)] = E(r, 0) \equiv 2\int_0^r f(r,t)dt\ .$$

Let us introduce formally

$$V(r) \equiv \tfrac{1}{2}\|E(r, \lambda)\|^2\ .$$

For $G \in \Omega_r$, the kernel of (1.4) generates the quadratic form

$$\|G\|^2 - \frac{\rho}{2}|G(0)|^2\ ,$$

and by Schwarz's inequality, exactly as in the proof of Proposition 1(c), this is positive

if and only if

$$\rho \leqslant \frac{2}{\max\limits_{G \in \Omega_r} \frac{|G(0)|^2}{\|G\|^2}} = \frac{2}{\|E(r,\lambda)\|^2} = \frac{1}{V(r)} .$$

Finally, if $r \leqslant s$, then $E(r,\lambda) \in \Omega_s$ and so, by (7.5),

$$[E(r,\lambda),E(s,\lambda)] = E(r,0) = E(\min(r,s),0) . \tag{7.6}$$

Thus

$$\begin{aligned} \frac{dV}{dr} &= \lim_{\delta \to 0} \frac{V(r)-V(r-\delta)}{\delta} = \tfrac{1}{2} \lim_{\delta \to 0} \frac{\|E(r,\lambda)\|^2 - \|E(r-\delta,\lambda)\|^2}{\delta} \\ &= \tfrac{1}{2} \lim_{\delta \to 0} \frac{[E(r,\lambda),E(r,\lambda) - E(r-\delta,\lambda)]}{\delta} . \end{aligned} \tag{7.7}$$

We evaluate this directly from the definition (7.1). To this end, as $f(r-\delta,t)$ vanishes for $|t| > r - \delta$, we have

$$(I+H_{r-\delta})f(r-\delta,s) = \begin{cases} 1, & |t| < r-\delta, \\ H_{r-\delta}f(r-\delta,s), & r-\delta < |t| < r. \end{cases}$$

Consequently by (1.1)

$$(I+H_r)f(r,s) - (I+H_{r-\delta})f(r-\delta,s) = \begin{cases} 0, & |t| < r - \delta, \\ 1-H_{r-\delta}f(r-\delta,s) , & r - \delta < |t| < r , \end{cases}$$

and from (7.7) and (7.1) we obtain

$$\frac{dV(r)}{dr} = \tfrac{1}{2} \lim_{\delta \to 0} \frac{1}{\delta} \int_{r-\delta<|t|<r} dt\, f(r,t) \left\{ 1-\tfrac{1}{2} \int_{-r+\delta}^{r-\delta} H(|t-s|)f(r-\delta,s)ds \right\} .$$

This limit evidently consists of the sum of values of the (smooth) integrand at the points $t = \pm r$, namely

$$\tfrac{1}{2} f(r,r)\left\{1-\tfrac{1}{2} \int_{-r}^{r} H(|r-s|)f(r,s)ds \right\} + \tfrac{1}{2} f(r,-r)\left\{1-\tfrac{1}{2} \int_{-r}^{r} H(r+s)f(r,s)ds \right\},$$

and from (1.1) this is

$$\frac{dV}{dr} = \tfrac{1}{2} f(r,r)f(r,r) + \tfrac{1}{2} f(r,-r)f(r,-r) = f^2(r,r) .$$

We thus obtain the relationship expressed by (1.2) and (1.3).

Let

$$A(r) \equiv f^2(r,r) .$$

To identify $A(r)$ with the area of a vocal tract, we turn to (1.5). Here, to parallel our reasoning of (4.8)-(4.11), we could start with the collection $\Omega^0_{r-\epsilon}$ of functions $G_{r-\epsilon}$ for which $\lambda^2 G_{r-\epsilon}(\lambda) \in \Omega_{r-\epsilon}$, and argue heuristically on the strength of (7.5) and (7.2) that

$$0 = [\lambda^2 G_{r-\epsilon}(\lambda),E(r,\lambda)] = [G_{r-\epsilon}(\lambda), \lambda^2 E(r,\lambda)] .$$

Likewise, since for $\eta < \epsilon$

$$\left[G_{r-\epsilon}(\lambda), \frac{E(r,\lambda) - E(r-\eta,\lambda)}{\eta}\right] = \frac{G_{r-\epsilon}(0) - G_{r-\epsilon}(0)}{\eta} = 0,$$

We would expect $\left[G_{r-\epsilon}, dE(r,\lambda)/dr\right] = 0$, and similarly $\left[G_{r-\epsilon}, d^2E(r,\lambda)/dr^2\right] = 0$. Thus we would picture $\lambda^2 E(r,\lambda), dE(r,\lambda)/dr$, and $d^2E(r,\lambda)/dr^2$ as orthogonal to $\Omega^0_{r-\epsilon}$. Now, of course, none of these three functions belongs to Ω_r, for each grows too fast, but we can form a linear combination in which the components lying outside Ω_r cancel; indeed as $dE(r,\lambda)/dr$ is the continuous analogue of $P_K(z)P_K(1)$, by analogy with (4.10) we can expect the relevant linear combination to be $\lambda^2 E(r,\lambda) + f^2(r,r)\,(d/dr)\,f^{-2}(r,r)\,dE/dr$. Since the union of $\Omega^0_{r-\epsilon}$ for $\epsilon > 0$ is easily seen to be dense in Ω_r, this last function, being orthogonal to each $\Omega^0_{r-\epsilon}$ must then vanish identically. Although this approach is direct and simple, technical obstacles arise since the scalar product is not even defined for functions like $\lambda^2 E(r,\lambda)$ which grow as rapidly as λ. We can, however, easily circumvent this difficulty by introducing a convergence factor.

Let us temporarily assume that H in (1.1) is continuously differentiable; then, arguing just as we did earlier, $f_{22}(r,t)$ and $f_{11}(r,t)$ are continuous functions of t. Let

$$Q_\epsilon(\lambda) = \left(\frac{\sin \lambda\epsilon/3}{\lambda\epsilon/3}\right)^3 .$$

Then $Q_\epsilon(\lambda)$ and $\lambda^2 Q_\epsilon(\lambda)$ are in Ω_ϵ, and $\lim_{\epsilon\to 0} Q_\epsilon(\lambda) = 1$, uniformly on compact sets. We will use Q_ϵ for technical purposes only, as a convergence factor. With $G_{r-2\epsilon} \in \Omega_{r-2\epsilon}$, the function $G_{r-2\epsilon}(\lambda)\lambda^2 Q_\epsilon(\lambda) \in \Omega_{r-\epsilon}$, and from (7.4) and (7.2) we find

$$0 = [G_{r-2\epsilon}(\lambda)\lambda^2 Q_\epsilon(\lambda),\ E(r,\lambda)] = [G_{r-2\epsilon}(\lambda)\ ,\ Q_\epsilon(\lambda)\lambda^2 E(r,\lambda)]\ . \tag{7.8}$$

Now $\lambda^2 E(r,\lambda)$ is not in Ω_r, not being square-integrable; indeed, integrating by parts in (7.4) we find that

$$\lambda^2 E(r,\lambda) = f_2(r,r)\ 2\cos r\lambda + f(r,r)\lambda\ 2\sin r\lambda + e_1(r,\lambda)\ , \tag{7.9}$$

with $e_1 \in \Omega_r$. Continuing, when $r - \eta > r - \epsilon$, by (7.5) and (7.2),

$$0 = \left[G_{r-2\epsilon}(\lambda)Q_\epsilon(\lambda), \frac{E(r,\lambda) - E(r-\eta,\lambda)}{\eta}\right] = \left[G_{r-2\epsilon}(\lambda), Q_\epsilon(\lambda)\,\frac{E(r,\lambda) - E(r-\eta,\lambda)}{\eta}\right]. \tag{7.10}$$

Now from (7.4)

$$\frac{dE(r,\lambda)}{dr} = f(r,r)\ 2\cos\lambda r + \int_{-r}^{r} f_1(r,t)e^{it\lambda}dt\ , \tag{7.11}$$

so that, since dE/dr grows no faster than a constant, $Q_\epsilon(\lambda)\,d\Omega(r,\lambda)/dr \in \Omega_{r+\epsilon}$. Again from (7.4), $\eta^{-1}\{E(r,\lambda) - E(r-\eta,\lambda)\}$ converges to $dE(r,\lambda)/dr$ pointwise and uniformly on compact sets in λ, so that for each $\epsilon > 0$

$$\lim_{\eta\to 0} Q_\epsilon(\lambda)\left\{\frac{E(r,\lambda) - E(r-\eta,\lambda)}{\eta}\right\} = Q_\epsilon(\lambda)\,\frac{dE(r,\lambda)}{dr}\ ,$$

in the metric of $L^2(-\infty,\infty)$, and we see from (7.3) and (7.10) that

$$\left[G_{r-2\epsilon}(\lambda),\, Q_\epsilon(\lambda)\,\frac{dE(r,\lambda)}{dr}\right] = 0\,.$$

By the same argument applied to $\eta^{-2}\{E(r+\eta,\lambda)-2E(r,\lambda)+E(r-\eta,\lambda)\}$ we find that also $Q_\epsilon(\lambda)\,d^2E(r,\lambda)/dr^2 \in \Omega_{r+\epsilon}$ is orthogonal to $\Omega_{r-2\epsilon}$. Now introducing

$$y(r,\lambda) \equiv f^{-2}(r,r)\,\frac{dE(r,\lambda)}{dr} = \frac{1}{A(r)}\,\frac{dE(r,\lambda)}{dr}$$

we see from (7.11) that

$$\frac{dy(r,\lambda)}{dr} = -\frac{f'(r,r)2\cos\lambda r}{f^2(r,r)} - \frac{\lambda\, 2\sin\lambda r}{f(r,r)} + \frac{f_1(r,r)2\cos\lambda r}{f^2(r,r)} + e_2(r,\lambda)\,,$$

with $e_2(r,\lambda) \in \Omega_r$. Set

$$S(r,\lambda) \equiv \lambda^2 E(r,\lambda) + A(r)\,\frac{dy(r,\lambda)}{dr}\,.$$

Since $f'(r,r) = f_1(r,r) + f_2(r,r)$, we see from (7.9) that the linear combination defining $S(r,\lambda)$ has eliminated the rapidly growing terms, so that $S(r,\lambda) \in \Omega_r$. Simultaneously, since $dy(r,\lambda)/dr$ is a linear combination of $dE(r,\lambda)/dr$ and $d^2E(r,\lambda)/dr^2$, $Q_\epsilon(\lambda)\{dy(r,\lambda)/dr\}$ is orthogonal to $\Omega_{r-2\epsilon}$. Combining this with (7.8) and (7.2), we obtain

$$0 = [G_{r-2\epsilon}(\lambda), Q_\epsilon(\lambda)S(r,\lambda)] = [G_{r-2\epsilon}(\lambda)Q_\epsilon(\lambda),\, S(r,\lambda)]\,. \tag{7.12}$$

Now as any function in $L^2(-r,r)$ can be approximated arbitrarily closely in the metric of L^2 by functions of $L^2(-r+2\epsilon,r-2\epsilon)$ with $\epsilon \to 0$, the collection $\{\Omega_{r-2\epsilon},\, \epsilon > 0\}$ is dense in Ω_r in the L^2 norm. Since $Q_\epsilon(\lambda) \to 1$ uniformly on compact subsets, and $|Q_\epsilon(\lambda)| \leqslant 1$, the same is true of the set of functions $\{G_{r-2\epsilon}(\lambda)Q_\epsilon(\lambda) | G \in \Omega_{r-2\epsilon}, \epsilon > 0\}$. From (7.3) we see that this set is dense in E_r also in the norm $\|\cdot\|$, so that by (7.12)

$$S(r,\lambda) \equiv 0\,. \tag{7.13}$$

If H is merely continuous, we approximate it uniformly on $[0,2a]$ by a sequence H_m of continuously differentiable functions, and generate corresponding functions $E_m(r,\lambda)$, $A_m(r)$, and $y_m(r,\lambda)$. From (1.1), evidently $E_m(r,\lambda) \to E(r,\lambda)$ for each λ, and $A_m(r) \to A(r)$ uniformly in $r \leqslant a$, so that by (7.13)

$$\frac{dy_m(r,\lambda)}{dr} \to -\frac{\lambda^2 E(r,\lambda)}{A(r)}\,.$$

We conclude that $y_m(r,\lambda)$ approaches $y(r,\lambda)$ and that the latter is differentiable in r, satisfying

$$\frac{dy(r,\lambda)}{dr} = -\frac{\lambda^2 E(r,\lambda)}{A(r)}\,.$$

As the right-hand side is continuously differentiable, so is dy/dr, and now another differentiation with respect to r yields (1.5). From the equation satisfied by $f_2(r,t)$, or more simply because $f(r,t)$ is even, $f_2(0,0) = 0$. Thus we see from (7.11), (7.9), and (7.13) that $y(r,\lambda)$ is the solution of (1.5) which satisfies the initial conditions $y(0,\lambda) = 1$, $y'(0,\lambda) = 0$.

REFERENCES

[1] K. P. BUBE and R. BURRIDGE, *The one-dimensional inverse problem of reflection seismology*, SIAM Review, 25 (1983), to appear.

[2] R. BURRIDGE, *The Gelfand-Levitan, the Marchenko, and the Gopinath-Sondhi integral equations of inverse scattering theory, regarded in the context of inverse impulse-response problems*, Wave Motion, 2 (1980), pp. 305-323.

[3] R. E. CAFLISCH, *An inverse problem for Toeplitz matrices and the synthesis of discrete transmission lines*, Linear Algebra and its Applications, 38 (1980), pp. 255-272.

[4] K. M. CASE and M. KAC, *A discrete version of the inverse scattering problem*, J. Math. Phys., 14 (1973), pp. 594-603.

[5] K. M. CASE, *Inverse scattering, orthogonal polynomials, and linear estimation*, Topics in Functional Analysis, I. Gohberg and M. Kac, eds., Academic Press, New York, 1978, pp. 25-43.

[6] B. GOPINATH and M. M. SONDHI, *Inversion of the telegraph equation and the synthesis of non-uniform lines*, Proc. IEEE, 59 (1971), pp. 383-392.

[7] T. KAILATH, A. VIEIRA, and M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Review, 20 (1978), pp. 106-119.

[8] M. G. KREIN, *On integral equations which generate second-order differential equations* (Russian); Doklady Akad. Nauk. SSSR, 97 (1954), pp. 21-24.

[9] ———, *On the Sturm-Liouville boundary value problem in the interval* $(0,\infty)$ *and a class of integral equations* (Russian); Doklady Akad. Nauk SSSR, 73 (1950), pp. 1125-1128.

[10] ———, *Solution of the inverse Sturm-Liouville problem* (Russian); Doklady Akad. Nauk SSSR, 76 (1951), pp. 21-24.

[11] H. J. LANDAU, *The classical moment problem: Hilbertian methods*, J. Functional Analysis, 38 (1980), pp. 255-272.

[12] M. M. SONDHI and B. GOPINATH, *Determination of vocal-tract shape from impulse response at the lips*, J. Acoust. Soc. Am., 49 (1971), pp. 1867-1873.

[13] M. M. SONDHI, *Two acoustical inverse problems in speech and hearing*, Symposium on Scattering Theory, Lecture Notes in Physics, 130, Springer-Verlag, New York, 1980, pp. 290-300.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0001 $01.25/0

THE MORSE LEMMA IN INFINITE DIMENSIONS VIA SINGULARITY THEORY*

MARTIN GOLUBITSKY† AND JERROLD MARSDEN‡

Abstract. An infinite dimensional Morse lemma is proved using the deformation lemma from singularity theory. It is shown that the versions of the Morse lemmas due to Palais and Tromba are special cases. An infinite dimensional splitting lemma is proved. The relationship of the work here to other approaches in the literature in discussed.

Introduction. This paper shows that when the singularity theory proof of the Morse lemma is extended to infinite dimensions, it gives a result better than the best available. The best available Morse lemma is that of Tromba [1976], [1981] which improves upon the usual Morse–Palais lemma (cf. Palais [1963], [1969]) for the following crucial reason: The Morse–Palais lemma assumes that the second derivative of the function at its critical point is strongly nondegenerate in the sense of defining an isomorphism between the space and its dual. Such a hypothesis is not satisfied in standard elliptic variational problems; however, the hypotheses of Tromba's Morse lemma are normally verified in such problems. Specific examples are presented in Buchner, Marsden and Schecter [1983]; for others see

(a) Tromba [1976], [1981] for geodesics and minimal surfaces;

(b) Choquet-Bruhat and Marsden [1976] and Arms, Marsden and Moncrief [1982] for general relativity;

(c) Ball, Knops and Marsden [1978] and Marsden and Hughes [1983] for elasticity.

In conjunction with the Morse lemma are questions of

1. normal forms for more degenerate singularities and

2. a splitting lemma and reduction to finite dimensional catastrophe theory.

Such questions have been studied by Magnus [1976], [1978], [1979], Arkeryd [1979] and Chillingworth [1980], but under hypotheses similar to those of the Morse–Palais lemma. In view of the difficulties with these hypotheses, it is important to also carry this program out under more applicable hypotheses. Such a setting is provided here. A related setting for a normal form theory in infinite dimensions is presented in Beeson and Tromba [1981]. Their situation is further complicated by the presence of a group action. A closely related setting is given in Dangelmayr [1979] and Magnus [1980].

The plan of the paper is as follows:

1. Theorem A in §2 gives conditions under which two given functions are related by a diffeomorphism in a neighborhood of a singular point.

2. Theorem B in §3 is the Morse–Tromba lemma and is shown to be a straightforward consequence of Theorem A.

3. Section 4 discusses the splitting lemma and the associated reduction to finite dimensional catastrophe theory.

Finally, we note that the ideas in Theorem A below are useful in the study of vector fields. In particular, the methods can be used to deal with some C^∞-flat ambiguities in normal forms of vector fields at a singular point. These topics will be the subject of other publications.

*Received by the editors June 28, 1982, and in revised form October 29, 1982.

†Department of Mathematics, University of Houston, Houston, Texas 77004.

‡Department of Mathematics, University of California, Berkeley, California 94720.

1. The singularity theory method. To put the methods in perspective, we shall recall some of the ideas of singularity theory with a view towards the Morse lemma. The basic methods of singularity theory under the notion of k-determining are contained in Mather [1970], Siersma [1974] and Wasserman [1974], though they are not stated there in precisely the form we use here.

One of the goals of singularity theory is to bring functions into normal form in a neighborhood of a singular point. The procedure for doing so involves two steps:

1. *The analytical step.* This step gives criteria for when two functions are related by a diffeomorphism. This is done using what is called the deformation method and involves the integration of ordinary differential equations.

2. *The algebraic step.* The verification of the hypotheses needed to guarantee that a function is related to a specific normal form by a diffeomorphism usually reduces to a problem in linear algebra.

Let us formalize these steps somewhat, with a view toward the Morse lemma in $\mathbb{R}^n$. Let g and h be smooth real valued functions defined on a neighborhood of the origin in $\mathbb{R}^n$ with $g(0)=h(0)=0$. We say that g and h are *right equivalent* if there is a C^∞ diffeomorphism ϕ defined on a neighborhood of 0 in $\mathbb{R}^n$ with $\phi(0)=0$ such that $g(x)=h(\phi(x))$. If $D\phi(0)=I=$identity, we say that g and h are *strongly right equivalent.*

The *Morse lemma in* $\mathbb{R}^n$ states that *if g is a C^∞ function satisfying $g(0)=0$ and $Dg(0)=0$ and if $D^2g(0)$ is a nondegenerate symmetric bilinear form of index k then g is strongly right equivalent to*

$$h(x)=-\left(x_1^2+\cdots+x_k^2\right)+x_{k+1}^2+\cdots+x_n^2.$$

The proof of the Morse lemma proceeds by two steps. First of all one shows that g is strongly right equivalent to $Q(x)=\frac{1}{2}D^2g(0)\cdot(x,x)$ by writing $g=Q+p$ and seeking a diffeomorphism that eliminates p. The method for doing this is described below. Once this is done, linear algebra is used to make a further coordinate change to diagonalize Q. In this paper we shall concentrate on the first step; in infinite dimensions the second step depends on having a suitable spectral theorem available.

Now we discuss the general procedure one uses to show that $Q+p$ is right equivalent to Q. Let $\mathcal{O}_g$ denote the set of functions which are right equivalent to g. Let Tg denote the formal tangent space to $\mathcal{O}_g$ at g. More precisely, let ϕ_t be a curve of diffeomorphism with $\phi_t(0)=0$ and $\phi_0=\mathrm{Id}$; then $\psi_t(x)=g(\phi_t(x))$ is a curve in $\mathcal{O}_g$ with $\psi_0=g$. It follows from the chain rule that

$$\left(\frac{d}{dt}\psi_t\right)\Big|_{t=0}(x)=Dg(x)\cdot A(x)$$

where $A(x)=\dot{\phi}_t(x)|_{t=0}$. Thus, a typical element of Tg has the form $Dg(x)\cdot A(x)$, where $A(0)=0$.

The *deformation lemma* is as follows:

Let $g_t=Q+tp$. Assume

(D1) $p\in TQ$,

(D2) $Tg_t=TQ$ for all $t\in[0,1]$.

Then $Q+p$ is right equivalent to Q.

For strong right equivalence one modifies Tg to include the condition $DA(0)=0$, corresponding to $D\phi(0)=I$. A complete proof of the deformation lemma in the infinite dimensional case is given in the next section.

Next one uses the nondegeneracy hypotheses to show that indeed (D1) and (D2) are satisfied. For the Morse lemma this can be done directly since DQ can be identified

with an invertible linear transformation T by $DQ(x)\cdot y=\langle x, Ty\rangle$. Then in order to write $p(x)$ in the form $DQ(x)\cdot A(x)$, one can use Taylor's theorem to write $p(x)=\langle x,\hat{p}(x)\rangle$ and let $A(x)=T^{-1}\hat{p}(x)$. This is how (D1) is checked. In fact, one has proved that if $p(x)$ vanishes at the origin, then p is in TQ. Thus $TQ=\{p(x)|p(0)=0\}$.

One verifies $(D2)$ by showing that $Tg_t=\{p(x)|p(0)=0\}$ either by repeating the construction above for Tg_t or (preferably) by using a bit of algebraic machinery such as Nakayama's lemma. For more complicated singularities, the use of Nakayama's lemma is a practical necessity.

2. The deformation lemma. Let E be a Banach space. Let g,h: $U\to\mathbb{R}$ be C^k maps $(k\geq 1)$ defined on a neighborhood U of $0\in E$ and satisfy $g(0)=h(0)=0$. For $1\leq l\leq k$, we shall say that g *is* C^l *right equivalent to* h at 0 if there is a C^l diffeomorphism ϕ: $V\to W$ of neighborhoods of 0 in E such that

$$\phi(0)=0 \quad \text{and} \quad g(x)=h(\phi(x)) \quad \text{for all } x\in V.$$

Furthermore, if $D\phi(0)=I$, the identity, we say that g is C^l *strongly right equivalent* to h at 0.

THEOREM A. *Let f and p be C^k real valued functions $(k\geq 1)$ defined on a neighborhood of* 0 *in E and satisfy $f(0)=p(0)=0$. Make these assumptions*:

$(E1)$ *There is an $l\geq 1$, and a C^l map A: $U\to E$ defined on a neighborhood of* 0 *in E such that*

$$A(0)=0 \quad \textit{and} \quad p(y)=-Df(y)\cdot A(y) \quad \textit{for } y\in U.$$

$(E2)$ *There is a C^l map R: $U\to L(E,E)$ (the bounded operators on E with the norm topology) such that $R(0)=0$ and*

$$Dp(y)=Df(y)\circ R(y).$$

Then $f+p$ is C^l right equivalent to f at 0. *Furthermore, if in* $(E1)$, $DA(0)=0$, *then $f+p$ is C^l strongly right equivalent to f at* 0.

Remarks. (a) Conditions $(E1)$ and $(E2)$ are precise versions of $(D1)$ and $(D2)$ in the previous section, with Q replaced by f. Since Tf consists of functions of the form $Df(y)\cdot A(y)$, clearly $(E1)$ is expressing the same ideas as $(D1)$. For $(E2)$, observe that

$$Dp(y)=Df(y)\circ R(y)$$

is equivalent to saying that for all $t\in[0,1]$,

$$Df(y)+tDp(y)=Df(y)+tDf(y)\circ R(y)$$

i.e., with $f_t=f+tp$,

$$Df_t(y)=Df(y)\circ L_t(y)$$

where

$$L_t(y)=I+tR(y).$$

Thus there is a linear map relating $Df_t(y)$ and $Df(y)$. Since $R(0)=0$, $L_t(y)$ is invertible on a neighborhood of zero. Thus $(E2)$ is expressing $(D2)$. In fact the condition $R(0)=0$ can be replaced by $\|R(y)\|<1$ or invertibility of L_t for the right equivalence conclusion.

(b) If f is homogeneous of degree κ, i.e. $f(ty)=t^\kappa f(y)$ for a positive integer κ, and $(E2)$ holds then $(E1)$ holds. Also, if $R(0)=0$ then $DA(0)=0$ (so one has strong right

equivalence). To see this use $(E2)$ to write

$$p(y)=\int_0^1 Dp(\tau y)\cdot y\,d\tau=\int_0^1 Df(\tau y)\circ R(\tau y)\cdot y\,d\tau$$
$$=\int_0^1 \tau^{\kappa-1}Df(y)\circ R(\tau y)\cdot y\,d\tau.$$

Thus, $(E1)$ holds with $A(y)=-\int_0^1 \tau^{\kappa-1}R(\tau y)\cdot y\,d\tau$. Note that $A(0)=0$ automatically and $R(0)=0$ implies $DA(0)=0$.

(c) Theorem A readily generalizes to the case in which E is replaced by a Banach manifold. Condition $(E1)$ has intrinsic meaning independent of charts if A is a vector field on E. Condition $(E2)$ also makes intrinsic sense if R is a section of the bundle over E whose fiber at $x\in E$ is the set of continuous linear maps of T_xE to itself; this is a standard vector bundle associated with a manifold.

Proof of Theorem A. We first show that $f_t=f+tp$ is C^l right equivalent to f for small t. We find a curve of diffeomorphisms $\phi_t(x)=\phi(x,t)$ such that $\phi_0=I$, $\phi_t(0)=0$ and $f_t(\phi_t(x))=f(x)$. To do this, we seek a vector field $A_t(x)=A(x,t)$ of class C^l in x and t such that $A_t(0)=0$ and

$$p(y)=-Df_t(y)\cdot A_t(y).$$

If A_t is found, we let ϕ_t be its evolution operator defined by $\phi_t(x)=A_t(\phi_t(x))$ and $\phi_0=I$. Then we have

$$\frac{d}{dt}f_t(\phi_t(x))=p(\phi_t(x))+Df_t(\phi_t(x))\cdot A_t(\phi_t(x))=0.$$

Thus we get $f_t(\phi_t(x))=f(x)$ as desired. Note that $A_t(0)$ implies $\phi_t(0)=0$ and $DA_t(0)=0$ implies $D\phi_t(0)=I$.

Note that ϕ_t can be defined (on some neighborhood of 0) for as long a t-interval as A_t is defined. Indeed this follows from the fact that $A_t(0)=0$ and the continuous dependence of the solution curves on initial data. To construct A_t, we use $(E2)$ to write

$$Dp(y)=Df(y)\circ R(y).$$

Therefore,

$$Df_t(y)=Df(y)+tDp(y)=Df(y)(I+tR(y)).$$

Using this and $(E1)$, the equation $Df_t(y)\cdot A_t(y)=-p(y)$ becomes

$$Df(y)(I+tR(y))\cdot A_t(y)=Df(y)\cdot A(y).$$

Since $R(0)=0$, $I+tR(y)$ is invertible for y in a neighborhood of 0 and $0\le t\le 1$, so we can take

$$A_t(y)=(I+tR(y))^{-1}A(y).$$ □

3. Tromba's Morse lemma. This section shows that Tromba's Morse lemma is a direct and natural consequence of Theorem A. The setting is as follows.

Let E be a Banach space. Let $\langle\cdot,\cdot\rangle$ be an inner product on E and B: $E\times E\to\mathbb{R}$ a continuous symmetric bilinear form. Assume

$$h\colon E\to\mathbb{R} \text{ is } C^k, \qquad k\ge 3,$$

and satisfies:

$$h(0)=0, \quad Dh(0)=0, \quad D^2h(0)(u,v)=B(u,v).$$

Consider the following conditions:

($T1$) *There is a linear isomorphism* T: $E\to E$ *such that*

$$B(u,v)=\langle Tu,v\rangle \text{ for all } u,v\in E \text{ (nondegeneracy)}.$$

($T2$) h *has a* C^{k-1} *gradient relative to* $\langle\cdot,\cdot\rangle$, ∇h: $U\to E$; *i.e.*,

$$\langle \nabla h(y),u\rangle = Dh(y)\cdot u.$$

Note. ($T1$) implies that T itself is symmetric (i.e. $\langle Tu,v\rangle=\langle u,Tv\rangle$), since B is symmetric.

THEOREM B. (Morse–Tromba lemma). *If* ($T1$) *and* ($T2$) *hold, then* h *is* C^{k-2} *strongly right equivalent to* $f(x)=\frac{1}{2}B(x,x)$ *at* 0.

Proof. Note that $Df(y)\cdot u=B(y,u)=\langle Ty,u\rangle$, so $\nabla f(y)=Ty$. By Taylor's theorem, write $h=f+p$ where p is C^k, has a gradient and $p(0)=0$, $Dp(0)=0$ and $D^2p(0)=0$.

Since f is quadratic, Remark (b) following Theorem A, shows that it suffices to show that ($E2$) holds with $R(0)=0$.

To do this, use the fact that p has a $\langle\cdot,\cdot\rangle$ gradient to write

$$\begin{aligned} Dp(y)\cdot u &= \langle \nabla p(y),u\rangle \\ &= \left\langle \int_0^1 D\nabla p(\tau y)\cdot y\,d\tau, u\right\rangle. \end{aligned}$$

However, differentiating $Dp(x)\cdot u=\langle \nabla p(x),u\rangle$ in x, we see that $D\nabla p(x)$ is symmetric. Thus

$$Dp(y)\cdot u=\left\langle y, \left(\int_0^1 D\nabla p(\tau y)\cdot y\,d\tau\right)\cdot u\right\rangle.$$

Hence we can take $R(y)=\int_0^1 D\nabla p(\tau y)\cdot y\,d\tau$.

Note that $R(0)=0$ so we have strong right equivalence, as required. □

Remarks. (a) There are versions of this theorem for which E is a Banach manifold. The main difference is to let $\langle\cdot,\cdot\rangle$ depend on the base point. These versions can also be derived from Theorem A. Such generalizations are called for in minimal surfaces (see Tromba [1981]) and in fluid mechanics where E is a coadjoint orbit (see Arnold [1978, App. 5] and Ebin and Marsden [1970]).

(b) *If the hypothesis* ($T1$) *of Theorem* B *holds and if* $R(y)$ *exists and has a* C^{k-2} *adjoint* $R(y)^*$ *i.e.*

$$\langle R(y)^*u,v\rangle=\langle u,R(y)v\rangle,$$

then ($T2$) *holds.*

Indeed, $Df(y)\circ R(y)=Dp(y)$, so

$$\langle Ty,R(y)\cdot u\rangle = Dp(y)\cdot u,$$

so

$$\nabla p(y)=R(y)^*\cdot Ty.$$

Thus, if $(T1)$ and $(E2)$ hold, the extra condition that Theorem B requires is the existence of $R(y)^* \cdot Ty$. The $R(y)$ chosen in the proof of Theorem B is symmetric, but even when $(T1)$ holds it is possible to satisfy Theorem A even when $R(y)^* \cdot Ty$ does not exist. An example is given in the following paper.

(c) If $(T1)$ holds for two different inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, then $(T2)$ will hold for $\langle \cdot, \cdot \rangle_1$ if and only if it does for $\langle \cdot, \cdot \rangle_2$. Indeed, using obvious notation, the two gradients are related by

$$\nabla_1 h(y) = T_2 T_1^{-1} \nabla_1 h(y).$$

Thus, *changing the weak metric does not aid in the existence of the gradient.* Likewise, allowing $\langle \cdot, \cdot \rangle$ to depend on the base point, but assuming $\langle u, v \rangle_x = \langle T_x u, v \rangle_0$ for an isomorphism T_x (perhaps chart dependent) does not help with the existence of the gradient.

In fact, $\langle \cdot, \cdot \rangle$ can be any continuous symmetric bilinear form, degenerate or not and the proof still goes through. One can always choose $\langle \cdot, \cdot \rangle = B(\cdot, \cdot)$ although it may be computationally convenient to make a different choice. However, for the splitting lemma considered below, where T need not be invertible, the choice of $\langle \cdot, \cdot \rangle$ really will affect whether or not the gradient exists.

(d) Assume $(T1)$ holds for $f(x) = \frac{1}{2} B(x,x)$ and let $p(x) = h(x) - f(x)$, where h satisfies the conditions preceeding $(T1)$. Then *conditions* $(E1)$ *and* $(E2)$ *hold if and only if there exists a* C^l *map* $\tilde{R}$: $U \to L(E,E)$ *defined on a neighborhood* U *of* 0 *in* E *such that* $\tilde{R}(0) = 0$ *and*

$$\text{(R)} \qquad Dp(y) \cdot u = \langle y, \tilde{R}(y) \cdot u \rangle \quad \textit{for all } y \in U \textit{ and } u \in E.$$

Indeed, if (R) holds, $(E1)$ and $(E2)$ are verified with $A(y) = -T^{-1} \int_0^1 \tilde{R}(\tau y) \cdot \tau y \, d\tau$ and $R(y) = T^{-1} \tilde{R}(y)$. Conversely $(E2)$ gives condition (R) with $\tilde{R}(y) = TR(y)$.

(e) If E is finite dimensional or admits a "C^k duality mapping", then Theorem B can be improved to $k \geq 2$ and the right equivalence is C^{k-1} (and C^k away from 0). In finite dimensions this result is due to Kuiper [1972] and in infinite dimensions to Tuan and Ang [1979]. The same result can be proved by the methods of this paper by using the Whitney properties of the remainder term in the form given by Tuan and Ang [1979].

4. The splitting lemma. We now briefly discuss the splitting lemma of Gromoll and Meyer [1969], which enables one to reduce infinite dimensional catastrophe theory to the finite dimensional case. As usual, we want hypotheses that will be applicable to elliptic variational problems—see the following paper for specific examples. We shall work in the context of Theorem B.

Let E be a Banach space and h: $E \to \mathbb{R}$ be C^k, $k \geq 3$ defined in a neighborhood U of 0. Suppose $h(0) = 0$ and write $D^2 h(0)(u,v) = B(u,v)$. Let $\langle \cdot, \cdot \rangle$ be an inner product on E and assume

$(S1)$ *there is a Fredholm operator* T: $E \to E$ *of index* 0 *such that* $B(u,v) = \langle Tu, v \rangle$ *for all* $u, v \in E$.

Since T is Fredholm of index 0, we can write $E = K \oplus L$, where $K = \ker T$ and $L = \operatorname{range} T$ (note that T is symmetric). Denote points in $K \oplus L$ as pairs (x,y). Also assume:

$(S2) = (T2)$ h *has a* C^{k-1} *partial gradient* $\nabla_y h$: $U \to L$ (i.e. $\langle \nabla_y h(u), v \rangle = Dh(u)v$ for all $v \in L$) and $\nabla_y h(0,0) = 0$.

SPLITTING LEMMA. *There is a change of coordinates* $\bar{x}=x$, $\bar{y}=\eta_x(y)=\eta(x,y)$ *such that in a neighborhood of* $(0,0)$, h *has the form*

$$h(\bar{x},\bar{y})=\tfrac{1}{2}D_y^2g(0,0)(\bar{y},\bar{y})+r(\bar{x}),$$

where $\eta(0,0)=0$, $D_x\eta(0,0)=0$, $D_y\eta(0,0)=I$ *and* $r(0)=0$, $Dr(0)=0$ *and* $D^2r(0)=0$.

Proof. Consider

$$\nabla_y h\colon U\to L.$$

Let P be the projection of $E\to L$ whose kernel is K, and note that $P\circ T$ restricted to L is an isomorphism of L to itself. Observe that $D_y\nabla_y h(0,0)\colon L\to L$ is the operator $P\circ T$. Thus, the implicit function theorem guarantees that the equation

$$\nabla_y h(x,y)=0, \quad \text{i.e.,} \quad D_y h(x,y)=0$$

uniquely defines a function $y=F(x)$ near $(0,0)$; $F(0)=0$ and $F'(0)=0$.

Let $g(x,y)=h(x,y+F(x))$ and note that

$$Dg(x,y)=Dh(x,y+F(x))\circ\begin{pmatrix} I & 0\\ F'(x) & I\end{pmatrix}.$$

An easy computation shows that $Dg(x,y)=0$ implies $y=0$ and $D_xg(x,0)=0$ if and only if $Dh(x,F(x))=0$.

Theorem B (with h depending on parameters) shows that there exists an x-dependent change of coordinates $y=\Phi_x(\hat{y})$ such that

$$g(x,\hat{y})=\tfrac{1}{2}D_y^2g(x,0)(\hat{y},\hat{y})+g(x,0).$$

Theorem A can now be applied in a similar way as in the proof of Theorem B to the first term $\frac{1}{2}D_y^2g(x,0)(\hat{y},\hat{y})$ showing that there is a further coordinate change $\tilde{y}=\Psi_x(\hat{y})$ depending parametrically on x, which transforms this term to $\frac{1}{2}D_y^2g(0,0)(\tilde{y},\tilde{y})$. Putting these coordinate changes together gives the result. □

Remarks. 1. Letting $m=\dim K$, this splitting lemma is a generalization of the usual form of the splitting lemma: *a function whose Hessian has corank m at a critical point can be decomposed as the sum of a 2-flat function of m-variables and a nondegenerate quadratic form in the remaining variables.* See, for example, Wasserman [1974, p. 137].

2. The proofs we give are for $k\geq 3$ and give coordinate changes of class $k-2$. However, as in remark (e) following Theorem B, this can be improved to $k\geq 2$ and coordinate changes of class C^{k-1} if the Banach space L admits a C^k duality mapping.

3. To find the critical points of h, it is enough to find the critical points of g restricted to the *finite* dimensional space $K=\ker T$.

4. To find a normal form for h, it is enough to find one for $g|K$. Note that the computation of g is not necessarily an easy matter as its definition depends on the implicitly defined function F. However, using implicit differentiation one can in principle calculate the Taylor expression of g to any given order.

Since K is finite dimensional, ordinary catastrophe theory can be used to classify generically what happens in these situations. If h depends on parameters, the splitting lemma is to be used in a parametric way; a specific example is worked out in the following paper (see Example 7). The general results of Chillingworth [1980] also generalize to the present context. Finally, we refer to Magnus [1980] for a splitting lemma under hypotheses similar to those described here.

REFERENCES

L. ARKERYD [1979], *Thom's theorem for Banach spaces*, J. London Math. Soc., 19, pp. 359–370.

J. ARMS, J. MARSDEN AND V. MONCRIEF [1982], *The structure of the space of solutions of Einstein's equatios* II: *Several Killing fields and the Einstein-Yang-Mills equations*, Ann. Phys., 144, pp. 81–106.

V. ARNOLD [1978], *Mathematical Methods of Classical Mechanics*, Springer, New York.

J. M. BALL, R.J, KNOPS AND J. E. MARSDEN [1978], *Two examples in nonlinear elasticity*, Lecture Notes in Mathematics 166, Springer, New York, pp. 41–49.

M. BEESON AND A. J. TROMBA [1981], *The cusp catastrophe of Thom in the bifurcation of minimal surfaces*, preprint #462, SFB, Bonn, Manus. Math. to appear.

M. BUCHNER, J. MARSDEN AND S. SCHECTER [1983], *Examples for the infinite dimensional Morse lemma*, this Journal, this issue, pp. 1045–1055.

D. CHILLINGWORTH [1980], *A global genericity theorem for bifurcations in variational problems*, J. Funct. Anal., 35, pp. 251–278.

Y. CHOQUET-BRUHAT AND J. MARSDEN [1976], *Solution of the local mass problem in general relativity*, Comm. Math. Phys. 51, pp. 283–296.

G. DANGELMAYR [1979], *Catastrophes and bifurcations in variational problems*, in Structural Stability in Physics, Guttinger and Elkemeier, eds., Springer-Verlag, New York.

D. EBIN AND J. MARSDEN [1970], *Groups of diffeomorphism and the motion of an incompressible fluid*, Ann. Math., 92, pp. 102–163.

D. GROMOLL AND W. MEYER [1969], *On differentiable functions with isolated critical points*, Topology, 8, pp. 361–369.

N. H. KUIPER [1972], *C^1-equivalence of functions near isolated critical points*, in Symposium on Infinite-Dimensional Topology, Annals of Math. Studies 69, R. D. Anderson, ed., Princeton Univ. Press, Princeton, NJ

R. J. MAGNUS [1976], *On universal unfoldings of certain real functions on a Banach space*, Math. Proc. Camb. Phil. Soc., 81, pp. 91–95.

_____ [1978], *Determining in a class of germs on a reflexive Banach space*, Proc. Camb. Phil. Soc., 84, pp. 293–302.

_____ [1979], *Universal unfoldings in Banach spaces: reduction and stability*, Math. Proc. Camb. Phil. Soc., 86, pp. 41–55.

_____ [1980], *A splitting lemma for nonreflexive Banach spaces*, Math. Scand. 46, pp. 118–128.

J. MARSDEN AND T. HUGHES [1983], *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ

J. MATHER [1970], *Notes on right equivalence*, Univ. of Warwick, preprint.

R. PALAIS [1963], *Morse theory on Hilbert manifolds*, Topology 2, pp. 299–340.

_____ [1969], *The Morse lemma on Banach spaces*, Bull. Amer. Math. Soc. 75, pp. 968–971.

D. SIERSMA [1974], *Classification and deformation of singularities*, Thesis, Amsterdam.

A. J. TROMBA [1976], *Almost-Riemannian structures on Banach manifolds, the Morse lemma and the Darboux theorem*, Canad. J. Math., 28, pp. 640–652.

A. J. TROMBA [1981], *A sufficient condition for a critical point of a functional to be a minimum and its application to Plateau's problem*, preprint #408 SFB, Bonn. Math. Ann. (to appear).

V. T. TUAN AND D. D. ANG [1979], *A representation theorem for differentiable functions*, Proc. Amer. Math. Soc., 75, pp. 343–350.

G. WASSERMAN [1974], *Stability of Unfoldings*, Lecture Notes in Mathematics 393, Springer, New York.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0002 $01.25/0

EXAMPLES FOR THE INFINITE DIMENSIONAL MORSE LEMMA*

MICHAEL BUCHNER†, JERROLD MARSDEN‡ AND STEPHEN SCHECTER§

Abstract. Examples are presented which show how to use the Morse lemma in specific infinite dimensional examples and what can go wrong if various hypotheses are dropped. One of the examples shows that the version of the Morse lemma using singularity theory can hold, yet the hypotheses of the Morse–Palais and Morse–Tromba lemmas fail. Another example shows how to obtain a concrete normal form in infinite dimensions using the splitting lemma and hypotheses related to those in the Morse–Tromba lemma. An example of Dancer is given which shows that for the validity of the Morse lemma in Hilbert space, some hypotheses on the higher order terms must be made in addition to smoothness, if the quadratic term is only weakly nondegenerate. A general conjecture along these lines is made.

Introduction. In this paper we discuss several examples relevant to the Morse lemma and singularity theory in infinite dimensions.

We begin with some historical comments on the various methods that have been used to prove the Morse lemma. The *original method of Morse* uses induction on the dimension of the space and does not, as given, apply to infinite dimensions. See Milnor [1963] for this proof. The *Palais method* was introduced in Palais [1963]. It is a modification of the original method that works in Hilbert space under the hypothesis of strong nondegeneracy of the quadratic term.

The *Moser–Weinstein method* is a variant of the singularity theory method described in Golubitsky and Marsden [1983] (this issue, pp. 1037–1044). It was adapted to the Morse lemma by Palais [1969]. Rather than directly join the quadratic part f to $f+p$ by $f+tp$, as in the preceding paper, one joins df to $df+dp$ by $df+t\,dp$. Palais' [1969] theorem states the following: *if E is a Banach space, h: $E\to\mathbb{R}$ is C^3, $Dh(0)=0$, and $D^2h(0)$, regarded as a map of E to E^*, is an isomorphism, then there is a C^1 diffeomorphism ϕ defined on a neighborhood of 0 in E such that*

$$\phi(0)=0, \qquad D\phi(0)=I(=\textit{identity}),$$

and

$$h(\phi(x))=h(0)+\frac{1}{2}D^2h(0)\cdot(x,x).$$

In Hilbert space this result reduces to that in Palais [1963]. We call the condition on $D^2h(0)$ *strong nondegeneracy*. If the map of E to E^* associated to $D^2h(0)$ is injective, we say $D^2h(0)$ is *weakly nondegenerate*.

The *Morse–Tromba lemma* was introduced in Tromba [1976]. It is motivated by the fact that in many elliptic variational problems one does not have strong nondegeneracy of the quadratic term. Rather, this is changed to weak nondegeneracy at the expense of putting special hypotheses on the nonlinear terms. The necessity of weak nondegeneracy occurred already for Hamiltonian systems in Marsden [1968]. Tromba's original proof was an adaptation of Palais' [1963] proof. A proof of the Morse–Tromba lemma using the Moser–Weinstein method was given in Choquet-Bruhat, Fischer and Marsden [1979]. The Morse–Tromba lemma is Theorem B of Golubitsky and Marsden [1983]. The *singularity theory method*, described in that paper, yields a result strictly stronger than Tromba's. Examples 5 and 6 below illustrate this.

*Received by the editors June 28, 1982, and in revised form October 29, 1982.
†Department of Mathematics, University of Maryland, College Park, Maryland 20742.
‡Department of Mathematics, University of California, Berkeley, California 94720.
§Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

For spaces admitting a duality map (such as Hilbert space or $W^{s,p}$ spaces with p even), the Morse–Tromba lemma is valid for C^2 functions with C^1 changes of coordinates (by Remark (e) following Theorem B of Golubitsky and Marsden [1983]). We do not know a C^2 counterexample if E is a general Banach space. We conjecture that there is not such an example.

The Morse–Tromba lemma suggests the question: can the Morse–Palais lemma be generalized *without* putting conditions on the higher order terms? We conjecture that the answer is no. More precisely,

CONJECTURE. *Let E be a Banach space and let B: $E\times E\to\mathbb{R}$ be a continuous symmetric bilinear map such that $x\mapsto B(x,\cdot)$ is not an isomorphism of E and E^*. Let $f(x)=\frac{1}{2}B(x,x)$. Then there is a C^3 map p: $E\to\mathbb{R}$ with $p(0)=0$, $Dp(0)=0$, and $D^2p(0)=0$ such that f and $f+p$ are not C^1 right equivalent.*

For E a Hilbert space, this conjecture has been verified by E. N. Dancer (private communication). His class of examples is presented below in Example 8.

In the examples that follow, the labels $(E1)$, $(E2)$, $(T1)$, $(T2)$, $(S1)$, $(S2)$, Theorem A and Theorem B refer to Golubitsky and Marsden [1983]. A couple of these examples are simple and well known but are included for completeness.

Example 1. This example shows that *nondegeneracy* of $D^2h(0)$ *in the sense of* $(T1)$ *is not sufficient for the validity of the Morse lemma.* Let $E=l_2$ and let h be the C^∞ function

$$h(x)=\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n}x_n^2-\frac{1}{3}\sum_{n=1}^{\infty}x_n^3.$$

Let $\langle x,y\rangle=\sum_{n=1}^{\infty}(1/n)x_ny_n$. Then $(T1)$ holds with $T=I$. However $(T2)$ fails, since the only possibility would be

$$\nabla h(x)_n=x_n-nx_n^2,\qquad n=1,2,\cdots$$

which is not defined on open sets in l_2. Indeed, the Morse lemma fails for this function. The quadratic term has no zeros other than the origin, yet h vanishes on the sequence $(0,0,\cdots,3/2n,0,\cdots)$, which approaches 0 in l_2. If the cubic term is changed to $\frac{1}{3}\sum_{n=1}^{\infty}(1/n)x_n^3$ then the gradient exists and the Morse–Tromba lemma applies.

Example 2. This example shows that *Tromba's hypotheses* $(T1)$ *and* $(T2)$, *but not those of the Morse–Palais lemma, can be expected to hold for many elliptic variational problems.* If Ω is a bounded region in $\mathbb{R}^n$ with smooth boundary, $W^{s,p}(\Omega,\mathbb{R}^m)$ denotes the Sobolev space of maps u: $\Omega\to\mathbb{R}^m$ whose derivatives up to order s are in L^p (see Friedman [1969], for example). For $p=2$ we write $W^{s,2}=H^s$. If $m=1$ we write $W^{s,p}(\Omega,\mathbb{R})=W^{s,p}(\Omega)$.

Let us begin with the one-dimensional case.

(a) Let $E=H^1([a,b])$. We define the function g: $E\to\mathbb{R}$ by

$$g(u)=\int_a^b[u(x)]^2dx+\int_a^b[u(x)]^3dx=f(u)+p(u).$$

Composition properties of Sobolev spaces (Palais [1968]) show that g is C^∞. Considered as a linear map $E\to E^*$, the bilinear map $D^2g(0)$ is $u\mapsto(v\mapsto 2\int_a^b uv)$. This map is injective but not surjective. For example the delta function $\delta_x(v)=v(x)$ for $a<x<b$ is in E^* but not in the image of $D^2g(0)$: $E\to E^*$. Thus the hypotheses of the Morse–Palais lemma do not hold. (If $\int_a^b[u(x)]^2dx$ is replaced by $\int_a^b[u(x)]^2dx+\int_a^b[u'(x)]^2dx$, then

the hypotheses of the Morse–Palais lemma *do* hold; this quadratic functional is similar to the functionals used in the variational approach to geodesics.)

On the other hand, let $\langle \cdot, \cdot \rangle$ be the L^2 inner product on H^1. Then the gradient $\nabla g(u)$ relative to $\langle \cdot, \cdot \rangle$ is given by $\nabla g(u) = 2u + 3u^2$, which is C^∞. Moreover, $D\nabla g(0) = 2I$. Consequently, Tromba's Morse lemma applies, so g can be transformed to the functional $\int_a^b [u(x)]^2 dx$.

In this example the transformation can be seen directly. Observe that $g(u)$ can be written as $g(u) = \int_a^b [u(x)(1+u(x))^{1/2}]^2 dx$. Now if ϕ: $(c,d) \subset \mathbb{R} \to \mathbb{R}$ is C^∞ then $u \mapsto \phi \circ u$ is C^∞ on $\{u \in H^1 | c < u(x) < d \text{ for all } x \in [a,b]\}$. Hence the map $u \mapsto u(1+u)^{1/2}$ is C^∞ on $\{u \in H^1 | -1 < u(x) < \infty \text{ for all } x \in [a,b]\}$, has derivative the identity at $u=0$, and hence is a local diffeomorphism.

Tromba's proof of his Morse lemma applied to this example also yields the transformation $u \mapsto u(1+u)^{1/2}$. So does the proof of Theorem A. For if one solves $p = -df \cdot A$ by $A(u) = -u^2/2$ and $dp = df \circ R$ by $R(u) \cdot v = 3uv/2$, one obtains $A(u) = -u^2/2$ and for $A_t(u)$ we get the expression $-[1+3tu/2]^{-1}u^2/2$. Note that $A(u) = -\int_0^1 \tau R(\tau u) \cdot u \, d\tau$, in agreement with Remark (b) following Theorem A of Golubitsky and Marsden [1983]. Integrating this vector field leads to the inverse of the transformation $u \mapsto u(1+u)^{1/2}$.

(b) We now sketch a typical multiple integral variational problem in higher dimensions. (Proofs rely on standard elliptic theory and Sobolev estimates, which are omitted here.) Let E be $W_\partial^{s,p}(\Omega)$, the $W^{s,p}$ functions which are zero on $\partial\Omega$, and let $s > n/p + 1$. Consider h: $E \to \mathbb{R}$ defined by

$$h(u) = \int_\Omega W(Du)\, dx + \int_\Omega K(u)\, dx,$$

where W is a smooth function of $\mathbb{R}^n$ to $\mathbb{R}$, K is a smooth function of $\mathbb{R}$ to $\mathbb{R}$, and $Du(x)$ is identified with a column vector or a point in $\mathbb{R}^n$. Suppose that

$$W(0) = 0, \quad DW(0) = 0, \quad K(0) = DK(0) = D^2K(0) = 0$$

and

$$D^2W(0) \cdot (\xi, \eta) \geq c\|\xi\|\|\eta\| \quad \text{for all } \xi, \eta \in \mathbb{R}^n, \quad \text{where } c > 0.$$

Standard Sobolev inequalities (cf. Palais [1968, Thm. 9.10]) show that h is a smooth function. Let $\langle \cdot, \cdot \rangle$ on E be given by

$$\langle u, v \rangle = \int_\Omega Du \cdot Dv \, dx.$$

Then

$$Dh(u) \cdot v = \int_\Omega DW(Du) \cdot Dv \, dx + \int_\Omega DK(u) \cdot v \, dx$$

and

$$D^2h(0) \cdot (u,v) = \int_\Omega (Du)^T M(Dv)\, dx$$

where $D^2W(0) \cdot (\xi, \eta) = \xi^T M \eta$ for an $n \times n$ positive definite matrix M. Then $(T1)$ holds for $(Tu)(x) = \Delta^{-1} \operatorname{div}(MDv)$, using the classical fact that Δ: $W_\partial^{s,p}(\Omega) \to W^{s-2,p}(\Omega)$ is an

isomorphism (Friedman [1969]). Also, T is an isomorphism on these spaces, for, as is readily checked, div(MDu) is elliptic with trivial kernel. ($T2$) holds with

$$\nabla h(u)=\Delta^{-1}\operatorname{div}[DW(Du)]-\Delta^{-1}DK(u).$$

For this example, again the hypotheses of the Morse–Tromba lemma hold, but those of the Morse–Palais lemma do not. Examples like this occur in minimal surfaces (see Tromba [1981]) and in elasticity (see Chillingworth, Marsden and Wan [1982] and Marsden and Hughes [1983]).

Example 3. (a) This example will show that *Theorem* A *is not limited to functions of the form quadratic + higher order* (*as the Morse–Tromba lemma is*). Let $E=H^1([a,b])$ and let

$$g(u)=\int_a^b[u(x)]^3dx+\int_a^b[u(x)]^4dx.$$

Let

$$f(u)=\int_a^b[u(x)]^3dx \quad \text{and} \quad p(u)=\int_a^b[u(x)]^4dx.$$

The equation $p(u)=-Df(u)\cdot A(u)$ can be solved by $A(u)=-u^2/3$, and $Dp(u)=Df(u)\circ R(u)$ is solved by $R(u)\cdot v=4uv/3$. Note that

$$A(u)=-\int_0^1\tau^2R(\tau u)\cdot u\,d\tau.$$

(See Remark (b) following Theorem A in Golubitsky and Marsden [1983].) Hence g can be transformed to f by a C^∞ transformation. This transformation, as in Example 2a, can be found directly by writing $g(u)=\int_a^b[u(x)(1+u(x))^{1/3}]^3dx$. Then $\phi(u)=u(1+u)^{1/3}$ is a suitable transformation.

An easy calculation shows that the diffeomorphism obtained by integrating the vector field $A_t(u)=-[1+4tu/3]^{-1}u^2/3$ is the inverse of $\phi(u)=u(1+u)^{1/3}$.

(b) Let $E=H^1([a,b])$ and let $g(u)=f(u)+p(u)$ where $f(u)=\int_a^b[u(x)]^3dx$ and $p(u)=\{\int_a^b[u(x)]^2dx\}^2$. Theorem A applies to g and shows that g can be transformed to f. Indeed, $p(u)=-Df(u)\cdot A(u)$ can be solved by $A(u)=-\frac{1}{3}\int_a^b[u(x)]^2dx$ and $Dp(u)=Df(u)\circ R(u)$ is solved by $R(u)\cdot v=\frac{4}{3}\int_a^b u(x)v(x)dx$ (both $A(u)$ and $R(u)\cdot v$ are constant functions). Note that $A(u)=\int_0^1\tau R(\tau u)\cdot u\,d\tau$. On the other hand there does not seem to be any explicit diffeomorphism that one could write down by inspection of g. The conjugating diffeomorphism is given by integrating

$$\frac{\partial\phi}{\partial t}(u,t)=\frac{-\frac{1}{3}\int_a^b[\phi(u,t)]^2dx}{1+\frac{4}{3}t\int_a^b\phi(u,t)\,dx}, \qquad \phi(u,0)=u$$

and setting $t=1$. It seems unlikely this could be solved explicitly in any simple fashion.

Example 4. We now give an example of *a function h which is* C^3, ($T1$) *holds,* ∇h *exists and is continuous but is not* C^1, *and yet the Morse lemma fails.*

Thus the hypothesis that ∇h be C^1 cannot be weakened to C^0 in the Morse–Tromba lemma.

Let $E=L^q([0,1])$ and let $\phi\colon\mathbb{R}\to\mathbb{R}$ be a C^∞ function such that $\phi'(\lambda)=1$, $-1\le\lambda\le1$ and $\phi'(\lambda)=0$ if $|\lambda|\ge2$. We assume ϕ is monotone increasing with $\phi=-M$ for $\lambda\le-2$

and $\phi = M$ for $\lambda \geq 2$. Let h: $E \to \mathbb{R}$ be given by

$$h(u) = \frac{1}{2}\int_0^1 [u(x)]^2 dx + \frac{1}{3}\int_0^1 \phi\big([u(x)]^3\big)\, dx = f(u) + p(u).$$

For $q \geq 2$, f is clearly C^∞. Let $\langle \cdot, \cdot \rangle$ be the L^2 inner product on $L^q([0,1])$; then $(T1)$ holds with $T = I$. We claim that if q is an integer, then p is C^{q-1} but not C^q, and ∇p exists, is continuous, but is not C^1. Thus with $q \geq 4$ we get a C^3 function. Let us indicate the proof of these facts for $q = 4$.

To prove that p is C^3, we let $\psi(\lambda) = \phi(\lambda^3)$. By Taylor's theorem,

$$\psi(\lambda) = \sum_{k=0}^{3} \psi^{(k)}(\lambda_0)\frac{(\lambda - \lambda_0)^k}{k!} + R(\lambda, \lambda_0)(\lambda - \lambda_0)^3$$

where $\lim_{\lambda \to \lambda_0} R(\lambda, \lambda_0) = 0$ and, from the definition of ϕ, $\psi^{(k)}$ and R are bounded smooth functions. Thus, suppressing the argument x, for u and u_0 continuous functions of x we have the identity

$$3p(u) = \sum_{k=0}^{3} \int_0^1 \psi^{(k)}(u_0)\frac{(u - u_0)^k}{k!}\, dx + \int_0^1 R(u, u_0)(u - u_0)^3 dx.$$

Since $\psi^{(k)}(\lambda_0)$ and $R(\lambda, \lambda_0)$ are bounded continuous, $\psi^{(k)}(u_0)$ (resp. $R(u, u_0)$) extends to a continuous mapping from L^4 (resp. $L^4 \times L^4$) to L^4. Using this fact and the Schwarz inequality, it follows that $p(u)$ depends continuously on $u \in L^4$, and each integral above depends continuously on $(u, u_0) \in L^4 \times L^4$. Thus the identity holds for all $(u, u_0) \in L^4 \times L^4$. Since $\psi^{(k)}(u_0)$ $(k = 0, 1, 2, 3)$ is bounded, $(v_1, \cdots, v_k) \mapsto \int_0^1 \psi^{(k)}(u_0) v_1 \cdots v_k\, dx$ is a bounded multilinear functional on L^4. Using the Schwarz inequality and the Lebesgue dominated convergence theorem we see that the mapping that associates to $u_0 \in L^4$ the k-multilinear functional $(v_1, \cdots, v_k) \mapsto \int_0^1 \psi^{(k)}(u_0) v_1 \cdots v_k\, dx$ is continuous from L^4 to the bounded multilinear functionals on L^4. Also, $\lim_{u \to u_0} R(u, u_0) = 0$. It follows from the converse to Taylor's theorem (Abraham and Robbin [1967]) that p is C^3.

Now an easy check shows that $\nabla p(u)$ exists and is given by

$$\nabla p(u) = \frac{\psi'(u)}{3}.$$

If this were C^1, its derivative would be

$$u \mapsto \left(v \to \frac{\psi''(u)v}{3} \equiv P(u) \cdot v \right).$$

Choose a number a such that $\psi''(a)/3 \neq 0$ and let

$$u_n = \begin{cases} a & \text{on } [0, 1/n], \\ 0 & \text{elsewhere} \end{cases}$$

and

$$v_n = \begin{cases} \sqrt[4]{n} & \text{on } [0, 1/n], \\ 0 & \text{elsewhere.} \end{cases}$$

Then one sees that $u_n \to 0$ in L^4, $\|v_n\| = 1$ in L^4, but $P(u_n) \cdot v_n \not\to 0$ in L^4. Since $P(0) = 0$, $\nabla p(u)$ is not C^1. (One sees in a similar way that p is not C^4.)

Finally, we note that h has a sequence of critical points u_n approaching the origin, namely

$$u_n=\begin{cases}-1 & \text{on } [0,1/n],\\ 0 & \text{on } (1/n,1].\end{cases}$$

Since this is not true for f, the Morse lemma cannot hold for h.

Example 5. We give an example to show that $(E1)$, $(E2)$ *and* $(T1)$ *can hold, without* $(T2)$ *holding. Thus, the Morse lemma is valid, Theorem* A *applies, but Tromba's Theorem* B *does not.*

Let $E=l_1$, the space of sequences x_n with $\Sigma_{n=1}^{\infty}|x_n|<\infty$. Let $h=f+p$ where

$$f(x)=\frac{1}{2}\sum_{n=1}^{\infty}x_n^2=\frac{1}{2}\langle x,x\rangle,$$

$\langle\cdot,\cdot\rangle$ being the usual l_2 inner product, and

$$p(x)=\left(\sum_{n=1}^{\infty}x_n\right)x_1^2+x_2^3+x_3^3+\cdots.$$

Since p is induced by a continuous trilinear map, h is C^{∞}. Also, $(T1)$ holds with $T=I$. $(E1)$ holds with

$$A(x)=-\left(\left(\sum_{n=1}^{\infty}x_n\right)x_1,\,x_2,\,x_3,\cdots\right)$$

and $(E2)$ holds with

$$R(y)\cdot u=\left(\left(\sum_{i=1}^{\infty}u_i\right)y_1+2\left(\sum_{i=1}^{\infty}y_i\right)u_1,\,3u_2y_2,\,3u_3y_3,\cdots\right)$$

as is easily checked. However $(T2)$ cannot hold using the l_2 inner product (or, by Remark (c), following Theorem B, any inner product such that $(T1)$ holds). If ∇h exists, so does ∇p (since $\nabla f(x)=x$). But ∇p would be

$$\nabla p(x)=\left(x_1^2+2x_1\left(\sum_{n=1}^{\infty}x_n\right),\,3x_2^2+x_1^2,\,3x_3^2+x_1^2,\cdots\right)$$

which is not in l_1. Note also that $R(y)$ does not have an everywhere defined l_2 adjoint; see Remark (b) following Theorem B.

Example 6. A variation on Examples 2 and 5 gives an example which is a prototype for problems in elasticity in which two bodies are in contact at a point. Like Example 5, this *example has* $(E1)$, $(E2)$ *and* $(T1)$ *holding, but not* $(T2)$.

Let $\Omega\subset\mathbb{R}^n$ be a region with smooth boundary and $0\in\Omega$; for instance, let Ω be the unit disk in the plane. Let $E=W_0^{s,p}$, $s>n/p+1$, the Sobolev space $W^{s,p}$ with Dirichlet boundary conditions, and let $h\colon E\to\mathbb{R}$ be given by

$$h(u)=\frac{1}{2}\int_{\Omega}\|Du\|^2dx+u(0)\int_{\Omega}u^2dx.$$

As above, h is C^∞. Let $\langle u,v\rangle$ on E be defined by $\langle u,v\rangle=\int_\Omega(Du\cdot Dv)\,dx$. Then $(T1)$ holds with $T=I$. Let

$$f(u)=\frac{1}{2}\int_\Omega\|Du\|^2dx \text{ and } p(u)=u(0)\int_\Omega u^2\,dx.$$

We show that p cannot have a gradient ∇p with respect to $\langle\cdot,\cdot\rangle$ (except in the case $s=1$, $n=1$), and thus $(T2)$ cannot hold (except in the case $s=1$, $n=1$, in which case $(T2)$ holds) for ∇p would have to satisfy

$$\int_\Omega Dv\cdot D(\nabla p(u))\,dx=\int_\Omega Dv\cdot Du\,dx+v(0)\int_\Omega u^2\,dx+2u(0)\int_\Omega uv\,dx.$$

This implies

$$\Delta(u-\nabla p(u))=\left(\int_\Omega u^2\right)\delta_0+2u(0)u \quad \text{(as distributions)}$$

where δ_0 is the Dirac delta function at the origin. But δ_0 is not in $W^{s-2,p}$ unless $s=1$ and $n=1$, in which case $\delta_0\in W^{-1,p}$. In this latter case $\nabla p(u)$ is given by the formula $u-(\int_\Omega u^2)\Delta^{-1}\delta_0-2u(0)\Delta^{-1}u$ (using the fact that Δ: $W^{1,p}_\partial\to W^{-1,p}$ is an isomorphism). Thus Tromba's hypotheses are satisfied only in the case $s=1$, $n=1$. In the case $s=1$, $n=1$, $p=2$, the Palais–Morse lemma hypotheses are also satisfied since $\langle\cdot,\cdot\rangle$ is the Hilbert space inner product for H^1_∂.

On the other hand, for *arbitrary* s and n (with $s>n/2+1$), $p=-df\cdot A$ is solved by $A(u)=u(0)\Delta^{-1}u$, and $dp=df\circ R$ is solved by $R(u)\cdot v=-2u(0)\Delta^{-1}v-v(0)\Delta^{-1}u$ (note that $A(u)=-\int_0^1\tau R(\tau u)\cdot u\,d\tau$), so Theorem A applies.

Example 7 below concerns the splitting lemma under hypotheses compatible with Tromba's Morse lemma. We shall use the splitting lemma from the previous paper for problems in which there is an additional parameter.

Example 7. As in Example 2, let $s>n/p+1$ and $E=W^{s,p}_\partial(\Omega)$, the $W^{s,p}$ space with Dirichlet boundary conditions. Assume that λ_0 is a simple eigenvalue of the Laplacian Δ on Ω and define h: $E\times\mathbb{R}\to\mathbb{R}$ by

$$h(u,\lambda)=\int_\Omega\left[\frac{1}{2}\|Du\|^2+\frac{1}{2}(\lambda_0+\lambda)u^2+G(u)\right]dx,$$

where $G(t)=t^3+$(higher order terms) is a C^∞ function from $\mathbb{R}$ to $\mathbb{R}$. We shall apply the splitting lemma to h and bring it to normal form. We find that

(a) $Dh(u,\lambda)\cdot(v,\mu)=\int_\Omega[Du\cdot Dv+(\lambda_0+\lambda)uv+\frac{1}{2}\mu u^2+G'(u)v]\,dx$,

(b) $D^2h(u,\lambda)\cdot((v,\mu),\ (w,\nu))=\int_\Omega\ [Dv\cdot Dw+(\lambda_0+\lambda)vw+G''(u)vw+\nu uv+\mu uw]\,dx$,

(c) $D^3h(u,\lambda)\cdot((v,\mu),(w,\nu),(y,\sigma))=\int_\Omega[G'''(u)vwy+\sigma vw+\nu yv+\mu yw]\,dx$.

Define $\langle v,w\rangle=\int_\Omega Dv\cdot Dw\,dx$. Then $D^2h(0,0)\cdot((v,\mu),\ (w,\nu))=\int_\Omega[Dv\cdot Dw+\lambda_0vw]\,dx=\langle Tv,w\rangle$ where $Tv=(I-\lambda_0\Delta^{-1})v$. Since Δ is elliptic, T is Fredholm of index 0. The null space of T is $N(T)=\langle u_0\rangle$, where u_0 is an eigenfunction of Δ for the eigenvalue λ_0. The range of T is $R(T)=\langle u_0\rangle^\perp$, the space of vectors in E that are L^2-orthogonal to u_0. Let P be projection onto $\langle u_0\rangle^\perp$. Write $u\in E$ as $u=\alpha u_0+\tilde u$, $\tilde u\in\langle u_0\rangle^\perp$.

Let $\nabla h(u,\lambda)=(I-(\lambda_0+\lambda)\Delta^{-1})u-\Delta^{-1}G'(u)$. Then $Dh(u,\lambda)\cdot(v,0)=\langle\nabla h(u,\lambda),v\rangle$, and $P\nabla h$ is what was called $\nabla_y h$ in the splitting lemma. Solving $P\nabla h(u,\lambda)=0$ using the implicit function theorem gives a function $\tilde u(\alpha,\lambda)$ such that

$P\nabla h(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda)=0$. Now $\tilde{u}(0,0)=0$, and $D\tilde{u}(0,0)=0$ because the kernel of $P\circ D\nabla h(0,0)$ is $\langle u_0\rangle\times\mathbb{R}$. Clearly $\tilde{u}(0,\lambda)=0$ for all λ, since $\nabla h(0,\lambda)=0$ for all λ.

Let v denote a typical element of $\langle u_0\rangle^{\perp}$. Let $k(\alpha,v,\lambda)=h(\alpha u_0+\tilde{u}(\alpha,\lambda)+v,\lambda)$, so that $Dk(\alpha,0,\lambda)\cdot(0,w,0)=0$ for all $w\in\langle u_0\rangle^{\perp}$. There is then an (α,λ)-dependent change of coordinates $v=\eta_{(\alpha,\lambda)}(\bar{v})$ with $\eta_{(\alpha,\lambda)}(0)=0$ and $D\eta_{(\alpha,\lambda)}(0)=I$, such that

$$k\big(\alpha,\eta_{(\alpha,\lambda)}(\bar{v}),\lambda\big)=k(\alpha,0,\lambda)+\frac{1}{2}D^2k(\alpha,0,\lambda)(\bar{v},\bar{v}).$$

To find a normal form for k (and hence h) it remains to find a normal form for $g(\alpha,\lambda)\equiv k(\alpha,0,\lambda)=h(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda)$. Now,

(a) $$Dg(\alpha,\lambda)\cdot(\beta,\mu)=Dh\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)\cdot\big(\beta u_0+D\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu),\mu\big),$$

(b) $$D^2g(\alpha,\lambda)\cdot(\beta,\mu)^2=D^2h\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)\cdot\big(\beta u_0+D\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu),\mu\big)^2$$
$$+Dh\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)\cdot D^2\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu)^2,$$

(c)
$$D^3g(\alpha,\lambda)\cdot(\beta,\mu)^3=D^3h\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)\cdot\big(\beta u_0+D\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu),\mu\big)^3$$
$$+3D^2h\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)$$
$$\cdot\Big[\big(\beta u_0+D\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu),\mu\big),\ D^2\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu)^2\Big]$$
$$+Dh\big(\alpha u_0+\tilde{u}(\alpha,\lambda),\lambda\big)\cdot D^3\tilde{u}(\alpha,\lambda)\cdot(\beta,\mu)^3.$$

Therefore

(i) $g(0,0)=h(0,0)=0$,
(ii) $Dg(0,0)=0$ because $Dh(0,0)=0$,
(iii) $D^2g(0,0)=0$ because

$$D^2h(0,0)\cdot(\beta u_0,\mu)^2=\beta^2D^2h(0,0)\cdot(u_0,0)^2=\beta^2\langle Tu_0,u_0\rangle=\beta^2\langle 0,u_0\rangle=0,$$

(iv) $D^3g(0,0)\cdot(\beta,\mu)^3=\beta^3\int_\Omega G'''(0)u_0^3\,dx+3\mu\beta^2\int_\Omega u_0^2\,dx$,

using the formula for D^3h; the terms involving D^2h and Dh give 0.

Assume (by normalizing) that $\int_\Omega u_0^2\,dx=1$ and *assume* $\int_\Omega u_0^3\,dx\neq 0$. Since $G'''(0)\neq 0$, we have

$$g(\alpha,\lambda)=\frac{1}{3!}D^3g(0,0)\cdot(\alpha,\lambda)^3+\cdots=k\alpha^3+\frac{1}{2}\lambda\alpha^2+\cdots,\qquad k\neq 0.$$

Let us multiply α and λ by constants to put this in the form

$$g(\alpha,\lambda)=\alpha^3+3\lambda\alpha^2+\cdots.$$

The higher order terms are divisible by α^2, since $g(0,\lambda)=h(0,\lambda)=0$ (recall $\tilde{u}(0,\lambda)=0$), and $g_\alpha(0,\lambda)=0$ because $Dh(0,\lambda)=0$. Let us put g into normal form using the ideas in Wasserman [1975]. First note that $g(\alpha,\lambda)$ has the form

$$g(\alpha,\lambda)=\alpha^3z(\alpha)+3\alpha^2\lambda q(\alpha,\lambda)$$

where $z(0)=1$ and $q(0,0)=1$. By the universal unfolding theorem for cubic singularities, there are functions $\beta(\alpha,\lambda)$, $\sigma(\lambda)$ and $\tau(\lambda)$ such that

$$g(\alpha,\lambda)=h(\beta(\alpha,\lambda),\sigma(\lambda),\tau(\lambda))$$

where

$$h(\beta,\sigma,\tau)=\beta^3+\sigma\beta+\tau$$

and

$$\beta(0,0)=0,\quad \beta_\alpha(0,0)>0,\quad \sigma(0)=0,\quad \tau(0)=0.$$

Using the chain rule in some straightforward calculations we find that

$$\sigma'(0)=0 \quad \text{and} \quad \sigma''(0)<0.$$

Thus, there is a further change of coordinates $\mu=\mu(\lambda)$ with $\mu(0)=0$, $\mu'(0)>0$, such that

$$g(\alpha,\lambda)=[\beta(\alpha,\lambda)]^3-3[\mu(\lambda)]^2\beta(\alpha,\lambda)+\tau(\lambda).$$

Since $g(0,\lambda)=0$, $\{[\beta(0,\lambda)]^2-3[\mu(\lambda)]^2\}\beta(0,\lambda)+\tau(\lambda)=0$; and since $g_\alpha(0,\lambda)=0$, $3\{[\beta(0,\lambda)]^2-[\mu(\lambda)]^2\}\beta_\alpha(0,\lambda)=0$. Since $\beta_\alpha(0,\lambda)\neq 0$, $\beta(0,\lambda)=\varepsilon\mu(\lambda)$ where $\varepsilon=\pm 1$. Hence $\tau(\lambda)=2\varepsilon[\mu(\lambda)]^3$, so $g=\beta^3-3\mu^2\beta+2\varepsilon\mu^3$. Letting $\gamma=\beta-\varepsilon\mu$, we get $g=\gamma^3+3\varepsilon\gamma^2\mu$. If we differentiate each side of this equation twice with respect to α and once with respect to λ, and set $(\alpha,\lambda)=(0,0)$, we find that $1=[\gamma_\alpha(0,0)]^2[\gamma_\lambda(0,0)+\varepsilon\mu'(0)]$. But $\gamma_\lambda(0,0)=\beta_\lambda(0,0)-\varepsilon\mu'(0)=0$ and $\mu'(0)>0$; therefore $\varepsilon=1$. Thus we obtain the normal form

$$g(\alpha,\lambda)=\gamma^3+3\mu\gamma^2$$

in the new coordinates (γ,μ). Hence there is a change of coordinates respecting the parameter such that the higher order terms can be eliminated.

Note that g is the potential function for a *transcritical bifurcation*: if we set $g_\alpha(\alpha,\lambda)=0$ we get

$$3\alpha^2+6\lambda\alpha+2\alpha(\cdots)=0.$$

The solution set is therefore the λ-axis and a curve tangent at $(0,0)$ to the line $\alpha+2\lambda=0$. The expression $g(\alpha,\lambda)=\gamma^3+3\mu\gamma^2$ puts the potential function for this bifurcation problem into normal form.

Normal forms for the equations $g_\alpha=0$ by coordinate changes respecting the parameter are found in Golubitsky and Schaeffer [1979]; see Marsden and Hughes [1983, Chap. 7] for simple proofs adequate for the present example. Golubitsky and Schaeffer point out that for many bifurcation problems, the equation $g_\alpha=0$ can be put into normal form by a coordinate change respecting the parameter, but the potential function g cannot.

Our approach to Example 7 should be compared with, for example, Chillingworth [1974] and Zeeman [1976], which consider a one-dimensional problem in which difficulties with the function spaces do not occur (i.e the energy norm is a complete Hilbert space norm) and for which the bifurcation parameter is not treated as distinguished. The example of Beeson and Tromba [1981] has the function-space complications of our example (i.e. the energy norm is not complete) but has additional complications due to a group action. However there is no distinguished bifurcation parameter.

Example 8 (E. N. Dancer). *This example proves the conjecture in the introduction for Hilbert spaces.* Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and let B: $H \times H \to \mathbb{R}$ be a continuous symmetric bilinear map. There is a bounded self-adjoint operator L: $H \to H$ such that $B(x,y) = \langle Lx, y \rangle$. Suppose that L is *not* an isomorphism. Let $f(x) = \frac{1}{2}B(x,x)$. We shall find a continuous homogeneous cubic polynomial p: $H \to \mathbb{R}$ such that f and $f+p$ are not C^1 right equivalent in any neighborhood of the origin. Thus any generalization of the Morse–Palais lemma in Hilbert space *must* place restrictions on the perturbation p.

Let $\sigma(L)$ denote the spectrum of L. Since L is not an isomorphism, $0 \in \sigma(L)$.

Case 1. $N(L) = \varnothing$. Then $\nabla f(x) = Lx \neq 0$ for $x \neq 0$. We shall find a continuous homogeneous cubic polynomial p: $H \to \mathbb{R}$ such that $\nabla(f+p)(x) = Lx + \nabla p(x) = 0$ at points x arbitrarily close to 0. Then f and $f+p$ cannot be C^1 right equivalent.

There exist $w_n \in \sigma(L)$, $w_n \neq 0$, such that $w_n \to 0$. For each n let I_n be a closed interval centered at w_n such that the I_n are disjoint and radius $(I_n) < |w_n|/2$. Let P_n be the orthogonal projection corresponding to I_n that is given by the spectral theorem, and let $H_n = P_n H$. The subspaces H_n are mutually orthogonal subspaces of H, invariant under L, and $L \mid H_n$ has spectrum lying in I_n.

Choose $e_n \in H_n$ such that $\|e_n\| = 1$. Then $\|Le_n - w_n e_n\| < |w_n|/2$. Let $z_n = Le_n - w_n e_n$. Decompose z_n as $z_n = \sigma_n e_n + \tau_n y_n$ where $\langle e_n, y_n \rangle = 0$ and $\|y_n\| = 1$. If z_n is a multiple of e_n, set $y_n = 0$, $\tau_n = 0$. Then $y_n \in H_n$ (by invariance of H_n under L) and $|\sigma_n|, |\tau_n| < |w_n|/2$. We conclude that $Le_n = \mu_n e_n + \tau_n y_n$ where $\mu_n = w_n + \sigma_n$. Thus $|w_n|/2 < |\mu_n| < 3|w_n|/2$.

Define p_n on span $\{e_n, y_n\}$ by $p_n(\alpha e_n + \beta y_n) = \alpha^3 + (3\tau_n/\mu_n)\alpha^2\beta$. Notice that $|3\tau_n/\mu_n| < (3|w_n|/2)/(|w_n|/2) = 3$. We find that $L(\gamma_n e_n) + \nabla p_n(\gamma_n e_n) = 0$ provided

$$3\gamma_n^2 + \gamma_n \mu_n = 0$$

and

$$\frac{3\tau_n}{\mu_n}\gamma_n^2 + \gamma_n \tau_n = 0,$$

i.e., provided $\gamma_n = -\mu_n/3$.

Finally, define p on H by $p = \Sigma p_n$. Since all the e_n's and y_n's are mutually orthogonal, p is a continuous cubic polynomial.

(*Proof.* $p(x) = T(x,x,x)$ where $T(u,v,w)$ is the symmetric trilinear map defined by

$$T(u,v,w) = \sum \langle u, e_n \rangle \langle v, e_n \rangle \langle w, e_n \rangle + \sum \frac{\tau_n}{\mu_n} \langle u, y_n \rangle \langle v, e_n \rangle \langle w, e_n \rangle$$
$$+ \sum \frac{\tau_n}{\mu_n} \langle u, e_n \rangle \langle v, y_n \rangle \langle w, e_n \rangle + \sum \frac{\tau_n}{\mu_n} \langle u, e_n \rangle \langle v, e_n \rangle \langle w, y_n \rangle.$$

Each of these four sums is a *bounded* trilinear map. For example, the second sum is estimated as follows:

$$\left| \sum \frac{\tau_n}{\mu_n} \langle u, y_n \rangle \langle v, e_n \rangle \langle w, e_n \rangle \right| < \sum |\langle u, y_n \rangle \langle v, e_n \rangle \langle w, e_n \rangle|$$
$$\leq \left[\sum \langle u, y_n \rangle^2 \right]^{1/2} \left[\sum \langle v, e_n \rangle^2 \langle w, e_n \rangle^2 \right]^{1/2}$$
$$\leq \left[\sum \langle u, y_n \rangle^2 \right]^{1/2} \cdot \left[\sum \langle v, e_n \rangle^2 \cdot \sum \langle w, e_n \rangle^2 \right]^{1/2} \leq \|u\| \|v\| \|w\|. \quad)$$

We have $L(\gamma_n e_n) + \nabla p(\gamma_n e_n) = 0$ where $\gamma_n \to 0$.

Case 2. $N(L)\neq\varnothing$. Let $\{e_n\}$ be an orthonormal basis for $N(L)$. Let $p(x)=\Sigma\langle x,e_n\rangle^3$. Then $\nabla f(x)=0$ for all $x\in N(L)$, but $\nabla(f+p)(x)\neq 0$ if $x\neq 0$. Thus f and $f+p$ are not C^1 right equivalent.

REFERENCES

M. BEESON AND A. J. TROMBA [1981], *The cusp catastrophe of Thom in the bifurcation of minimal surfaces*, preprint #462 SFB, Bonn, Manus. Math., to appear.

D. R. J. CHILLINGWORTH [1974], *The catastrophe of a buckling beam*, Lecture Notes in Mathematics 468, Springer-Verlag, New York, pp. 86–91.

D. R. J. CHILLINGWORTH, J. E. MARSDEN AND Y. H. WAN [1982], *Symmetry and bifurcation in three-dimensional elasticity*, Arch. Rat. Mech. Anal., 80, pp. 295–331.

Y. CHOQUET-BRUHAT, A. FISCHER AND J. MARSDEN [1979], *Maximal hypersurfaces and positivity of mass*, in Isolated Gravitating Systems in General Relativity, J. Ehlers, ed., Proc. 1976 Varena conference, Italian Physical Society, pp. 322–395.

A. FRIEDMAN [1969], *Partial Differential Equations*, Holt, Rinehart, and Winston, New York.

M. GOLUBITSKY AND J. MARSDEN [1983], *The Morse lemma in infinite dimensions via singularity theory*, this Journal, this issue, pp. 1037–1044.

M. GOLUBITSKY AND D. SCHAEFFER [1979], *A theory for imperfect bifurcation via singularity theory*, Comm. Pure. Appl. Math., 32, pp. 21–98.

J. MARSDEN [1968], *Hamiltonian one parameter groups*, Arch. Rat. Mech. Anal., 28, pp. 323–361.

J. MARSDEN AND T. HUGHES [1983], *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ.

J. MILNOR [1963], *Morse Theory*, Princeton Univ. Press, Princeton, NJ.

R. PALAIS [1963], *Morse theory on Hilbert manifolds*, Topology 2, pp. 299–340.

_____ [1968], *Foundations of Global Non-linear Analysis*, Benjamin, Menlo Park, CA.

_____ [1969], *The Morse lemma on Banach spaces*, Bull. Amer. Math. Soc., 75, pp. 968–971.

A. J. TROMBA [1976], *Almost Riemannian structures on Banach manifolds, the Morse lemma and the Darboux theorem*, Canad. J. Math., 28, pp. 640–652.

_____ [1981], *A sufficient condition for a critical point of a functional to be a minimum and its application to Plateau's problem*, preprint #408, SFB Bonn, Math. Ann., to appear.

G. WASSERMAN [1975], *Stability of unfoldings in space and time*, Acta Math., 135, pp. 57–128.

C. ZEEMAN [1976], *Euler buckling*, Lecture Notes in Mathematics 525, Springer-Verlag, New York, pp. 373–395.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
00036-1410/83/1406-0003 $01.25/0

THE KORTEWEG–DE VRIES EQUATION, POSED IN A QUARTER-PLANE*

JERRY BONA† AND RAGNAR WINTHER‡

Abstract. An initial- and boundary-value problem for the Korteweg–de Vries equation is shown to be well-posed. The considered problem may serve as a model for unidirectional propagation of plane waves generated by a wavemaker in a uniform medium. Such models apply in regimes in which nonlinear and dispersive effects are of comparable small order.

AMS-MOS subject classification (1980). Primary 35B45, 35B65, 35C15, 35Q20, 45G10, 76B15

Key words. Korteweg–de Vries equation, surface water waves, existence, uniqueness and regularity, initial- and boundary-value problems, waves generated by a wavemaker

1. Introduction. The Korteweg–de Vries equation, originally suggested in connection with a certain regime of surface water waves, has been derived as a model for unidirectional propagation of small-amplitude long waves in a number of physical systems. Because of the range of its potential application, and because of its very interesting mathematical properties, this equation has been the object of prolific study in the last few years. These studies have generally concentrated on aspects of the pure initial-value problem,

$$u_t+u_x+uu_x+u_{xxx}=0, \tag{1.1}$$

$$u(x,0)=f(x), \tag{1.2}$$

for $x\in\mathbb{R}$ and $t\geq 0$, say. Equation (1.1) is a version of the Korteweg–de Vries equation in which the dependent and independent variables are nondimensional, but unscaled. The initial data f in (1.2) typically decays to zero at infinity, or is taken to be a periodic function, though these do not exhaust the theory thus far existent (cf. Bona and Schonbek [7] and Menikoff [20]). For comprehensive descriptions of results pertaining to the KdV equation, as (1.1) will be named subsequently, the reader may consult the review articles of Benjamin [3], Jeffrey and Kakutani [14], Lax [17], Miura [21], [22] and Scott, Chu and McLaughlin [24].

The applicability of the KdV equation in a particular context depends on many factors. Among the more universal of these is that the waves be unidirectional and essentially one-dimensional in character. It must generally be the case that, at least locally, the nonlinear and dispersive terms, uu_x and u_{xxx}, respectively, represent small corrections to the basic one-way propagator $u_t+u_x=0$ (cf. [4, §2]). In attempting to assess the performance of the KdV equation as a model for waves in a particular system, the pure initial-value problem may not be particularly convenient. There might be difficulty associated with determining the entire wave profile accurately at a given instant of time. One method of obtaining unidirectional waves to test the appurtenance of KdV is to generate waves at one end of a homogeneous stretch of the medium in question and to allow them to propagate into the initially undisturbed medium beyond the wavemaker. (Figure 1 shows an instance of this situation in the case of surface waves in a channel. For this system x is proportional to distance along the channel, t is

*Received by the editors August 15, 1981.
†Department of Mathematics, The University of Chicago, Chicago, Illinois 60637.
‡Institute for Informatics, University of Oslo, Oslo 3, Norway.

proportional to elapsed time and the dependent variable η is the deviation of the liquid's surface from its equilibrium position at the point x at time t. Here the dependent variable has been denoted η since u is usually reserved for a velocity in fluid flow contexts.) During the time when the waves propagate freely, it may be expected that KdV can apply. Of course any real medium will have finite extent, and once the waves have been influenced by another boundary the experiment should cease, as far as KdV is concerned. In such an experiment it may be comparatively easy to measure the passage of the generated waves at a fixed location at or away from the wavemaker. If this is the case, the generated waves can be determined, at or near the wavemaker, and at another station further away from the wavemaker. One could imagine using the measurement nearest the wavemaker as data for the KdV equation. It may then be possible to predict, perhaps numerically, the behavior of the waves further from the wavemaker on the basis of the KdV equation, and to compare the prediction with the measurements made well away from the wavemaker.

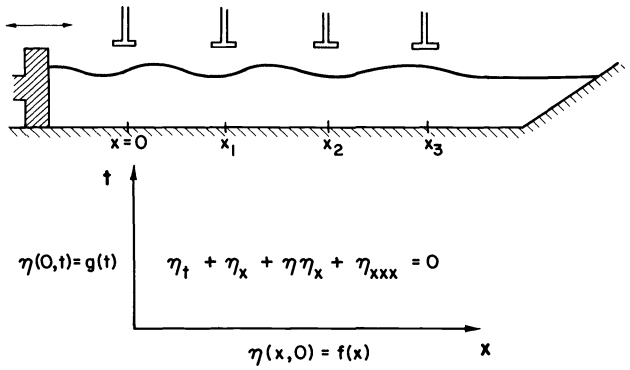


FIG. 1. *Sketch of the experimental configuration and the proposed mathematical model.*

The major accomplishment of the theory presented here is the demonstration that the program, just described, can, in principle, be carried out. Let us agree to fix the zero of the spatial coordinate x, which is along the direction of propagation, at the station nearest the wavemaker where a measurement is to be taken. Then the mathematical problem that accompanies the above discussion is expressed as the following initial- and boundary-value problem (cf. again Fig. 1).

$$
\begin{aligned}
&u_t+u_x+uu_x+u_{xxx}=0 && \text{for } x,t\geq 0,\\
&u(x,0)=f(x) && \text{for } x\geq 0,\\
&u(0,t)=g(t) && \text{for } t\geq 0.
\end{aligned} \tag{1.3}
$$

According to the above general discussion, it could be warranted to take $f\equiv 0$ and to assume that g, which is determined experimentally, is consistent with small-amplitude long-wavelength waves. These assumptions will play no role in the theory developed here.

All that will be required is that f and g exhibit smoothness, which is entirely appropriate to the use of KdV as a model equation, and that f decay to zero at infinity appropriately. The smoothness requirement extends to the origin, and results in a certain compatibility that must be satisfied between f and g. These conditions will be spelled out presently.

The same initial- and boundary-value problem has been analyzed for the alternative equation, proposed by Peregrine [23] and Benjamin et al. [4],

$$u_t+u_x+uu_x-u_{xxt}=0, \tag{1.4}$$

in [5]. Results related to those established in the latter reference will be derived and used in the attack on (1.3). The connection between KdV and (1.4) is a regularized version of problem (1.3), namely,

$$\begin{aligned} &u_t+u_x+uu_x+u_{xxx}-\varepsilon u_{xxt}=0 && \text{for } x,t\geq 0,\\ &u(x,0)=f(x) && \text{for } x\geq 0,\\ &u(0,t)=g(t) && \text{for } t\geq 0, \end{aligned} \tag{1.5}$$

where $\varepsilon>0$. The regularized problem (1.5) intervenes in a substantial way in the existence theory for (1.3) developed here. The regularized differential equation appearing in (1.5) is the same tool used already in [7] and [8] in discussions of various pure initial-value problems for KdV. The general outline of the theory herein is patterned after that developed in [8]. The technical difficulties presented by the nonhomogeneous boundary condition $u(0,t)=g(t)$, for $t\geq 0$, require a more delicate analysis than that effected in the last-quoted reference.

The present theory may be considered an extension of the earlier work of Ton [27] and Bona and Heard [6]. Ton's paper undertook the study of the problem,

$$\begin{aligned} &u_t+uu_x\pm u_{xxx}=0, && x,t>0,\\ &u(x,0)=f(x), && x\geq 0,\\ &u(0,t)=0, && t\geq 0. \end{aligned} \tag{1.6}$$

If the minus sign appears in front of the dispersive term, then the extra boundary condition $u_x(0,t)=0$, for $t>0$, is appended. For problem (1.6), with the positive sign taken, the methods exemplified in Lions' text [18], combined with the regularization used by Temam [26] in an early paper on the periodic initial-value problem for KdV, are used to obtain global existence of weak solutions and local existence of classical solutions. (The interval of existence is proportional to the inverse of $\|f\|_6$, in the notation to be introduced in §2.)

Actually, problem (1.6) is not an appropriate model for water waves in a uniform channel, as is suggested in [27]. For the differential equation in (1.6) is written in travelling coordinates, and consequently the boundary condition, if it is to correspond to observations of the disturbance at a fixed position in the channel, should be applied, not at $(0,t)$, for $t\geq 0$, but rather at $(-t,t)$, for $t\geq 0$. This awkwardness is easily obfuscated by the inclusion of the extra linear term u_x in the differential equation, an addition without serious consequence as regards Ton's mathematical proofs. A more serious objection to the theory developed in [27] is that the homogeneous boundary condition $u(0,t)=0$, for $t\geq 0$, is not well-suited to model waves generated by a wavemaker at one end of a uniform stretch of medium, as already explained. Moreover, for problems of long-wave propagation, it is not anticipated that the flow will develop singularities, and consequently it is expected that the model equation should have a global theory of classical solutions, corresponding to suitably smooth data. These drawbacks in the earlier theory are here shown to be methodological, and not inherently a property of the model equation.

In [6], a local existence theory for (1.3) is developed, using the methods of Kato [16]. The boundary data is required to be mildly smooth, but otherwise arbitrary. For

technical reasons, this theory has not, thus far, yielded solutions defined globally in time.

It is worth drawing attention to several comparisons which have been made with experimentally obtained data, pertaining to the originally conceived application of the KdV equation to small-amplitude surface water waves. We cite the studies of Zabusky and Galvin [31] and Hammack and Segur [13], and of Hammack [12] using (1.4). These studies all used pure initial-value problems for their theoretical predictions, even though the experimental configuration was exactly as described earlier in justifying the further study of the initial- and boundary-value problem considered here. That is, a uniform channel of water, initially at rest, had waves generated at one end by a wavemaker. The waves propagated down the channel and their passage was recorded at various stations along the channel. Entailed in each of these studies was a transformation of data measured over time, at a fixed location, to data measured spatially at a fixed instant of time. The approximate transformations used in the above-quoted studies introduce errors, which can be analyzed. In fact, the forthcoming work [10] addresses this issue in some detail, and consequently it is not taken up here, except to report that quite significant errors, particularly as regards the phase speed, can be expected when using the approach of converting the boundary-value problem to a pure initial-value problem.

It is also worth noting that, at least for surface water waves, damping effects need to be considered. Such effects were introduced, in an ad hoc way, in [12] and [13], and more systematically in [10]. An additional term that models the damping due to the boundary layers on the bottom and sides of a uniform channel of shallow water, at the level of approximation entailed in the KdV equation, has been derived carefully by Kakutani and Matsuuchi [15]. The incorporation of such dissipative terms in the initial- and boundary-value problem (1.3) is under study, but will not be addressed here.

The paper is organized as follows. Section two sets out the notation and terminology to be used subsequently and presents a sample of the main results in the paper. In §3 the regularized problem (1.5) is considered, and is shown to admit a satisfactory theory when ε is fixed and positive. A priori ε-independent bounds for solutions of the regularized problem are derived in §§4 and 5. Passage to the limit $\varepsilon \downarrow 0$ is effected in §6, where smooth solutions of the initial- and boundary-value problem (1.3) are shown to exist. The paper concludes with some commentary concerning aspects not covered in the present study.

2. Preliminaries and statement of the main result. For an arbitrary Banach space X, the associated norm will be denoted $\|\cdot\|_X$. The following spaces will intervene in the subsequent analysis.

If Ω is a bounded domain in $\mathbb{R}^n$, then $C^j(\bar{\Omega})$ denotes the space of real-valued functions which have classical derivatives up to order j in Ω, and whose derivatives, up to order j, extend to a continuous function on $\bar{\Omega}$. If $j=0$, $C^0(\bar{\Omega})$ will be denoted simply $C(\bar{\Omega})$. The norm on $C(\bar{\Omega})$ is

$$\|f\|_{C(\bar{\Omega})} = \sup_{x\in\Omega} |f(x)|,$$

and the norm on $C^j(\bar{\Omega})$ is

$$\|f\|_{C^j(\bar{\Omega})} = \sum_{|\alpha|\le j} \|\partial^\alpha f\|_{C(\bar{\Omega})}, \tag{2.1}$$

where $\alpha=(\alpha_1,\cdots,\alpha_n)$ is a multi-index of nonnegative integers, $|\alpha|=\alpha_1+\cdots+\alpha_n$, and

$$\partial^\alpha f(x) = \frac{\partial^{|\alpha|} f(x)}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}.$$

The notation ∂_x^r for $\partial^r/\partial x^r$ and ∂_t^r for $\partial^r/\partial t^r$ will be employed throughout when it is convenient. If Ω is unbounded, $C_b^j(\bar{\Omega})$ is defined exactly as in the case that Ω is bounded except that the function and its derivatives are required to be bounded. The norm is again defined by (2.1).

The space $C^\infty(\bar{\Omega}) = \cap_j C^j(\bar{\Omega})$ will be used, but its usual Fréchet-space topology will not be needed. $\mathcal{D}(\Omega)$ is the subspace of $C^\infty(\bar{\Omega})$ of functions with compact support in Ω. Its dual space, $\mathcal{D}'(\Omega)$, is the space of Schwartz distributions on Ω.

If Ω is open in $\mathbb{R}^n$, then $C^j(\Omega)$ is the continuous real-valued functions defined on Ω and possessing classical derivatives up to order j which are continuous on Ω. No restrictions are placed on the behavior of the functions near the boundary of Ω. This class can also be given a natural Fréchet-space topology, but this topology will not figure in the developments here. Naturally, $C^\infty(\Omega) = \cap_j C^j(\Omega)$.

If $T>0$, we will systematically use the abbreviation $C(0,T)$ for $C([0,T])$. Similarly, $C^m(0,T)$ will stand for $C^m([0,T])$.

For any real p in the range $[1,\infty)$, $L^p(\Omega)$ denotes the collection of real-valued Lebesgue measurable pth-power absolutely integrable functions defined on Ω. As usual, $L^\infty(\Omega)$ denotes the essentially bounded real-valued functions defined on Ω. These spaces get their usual norms,

$$\|f\|_{L^p(\Omega)} = \left\{ \int_\Omega |f(x)|^p dx \right\}^{1/p},$$

for $1 \le p < \infty$, and

$$\|f\|_{L^\infty(\Omega)} = \operatorname*{essential\,supremum}_{x\in\Omega} |f(x)|.$$

If $1 \le p \le \infty$, and $m \ge 0$ is an integer, let $W^{m,p}(\Omega)$ be the Sobolev space of $L^p(\Omega)$-functions whose distributional derivatives up to order m also lie in $L^p(\Omega)$. The norm on $W^{m,p}(\Omega)$ is

$$\|f\|_{W^{m,p}(\Omega)}^p = \sum_{|\alpha|\le m} \|\partial^\alpha f\|_{L^p(\Omega)}^p.$$

When $p=2$, $W^{m,p}(\Omega)$ will be denoted $H^m(\Omega)$. This is a Hilbert space, and $H^0(\Omega) = L^2(\Omega)$. For $s>0$, not necessarily an integer, $H^s(\Omega)$ is defined by interpolation. For $s>0$, $H_0^s(\Omega)$ is the closure in $H^s(\Omega)$ of $\mathcal{D}(\Omega)$. For $s>0$, $H^{-s}(\Omega)$ is the dual of $H_0^s(\Omega)$ with respect to the pairing which is the extension by continuity of the usual $L^2(\Omega)$-inner product. The noninteger-order Sobolev spaces only intrude at one point in our analysis, and then only in the interest of sharpness. Details concerning these spaces may be found in Lions and Magenes' work [19] or in Stein's text [25], for example. The notation $H^\infty(\Omega) = \cap_j H^j(\Omega)$ will be used for the C^∞-functions on Ω, all of whose derivatives lie in $L^2(\Omega)$.

Finally, $H_{\text{loc}}^s(\Omega)$ is the set of real-valued functions f defined on Ω such that, for each $\varphi \in \mathcal{D}(\Omega)$, $\varphi f \in H^s(\Omega)$. This space is equipped with the weakest topology such that all of the mappings $f \to \varphi f$, for $\varphi \in \mathcal{D}(\Omega)$, are continuous from $H_{\text{loc}}^s(\Omega)$ into $H^s(\Omega)$. With this topology, $H_{\text{loc}}^s(\Omega)$ is a Fréchet space (cf. Treves [28]). Let $\mathbb{R}^+$ denote the positive real numbers, $(0,\infty)$. A simple but pertinent example of the localized Sobolev spaces is $H_{\text{loc}}^s(\mathbb{R}^+)$. Interpreting the foregoing definitions in this special case, $g \in H_{\text{loc}}^s(\mathbb{R}^+)$ if and only if $g \in H^s(0,T)$, for all finite $T>0$. Moreover, $g_n \to g$ in $H_{\text{loc}}^s(\mathbb{R}^+)$ if and only if $g_n \to g$ in $H^s(0,T)$, for each $T>0$. Here and below, the abbreviation $H^s(0,T)$ has been used for $H^s((0,T))$.

In the analysis of the quarter-plane problem (1.3), the spaces $H^s(\Omega)$ will occur often, with s a positive integer and $\Omega=\mathbb{R}^+$ or $\Omega=(0,T)$. Because of their frequent occurrence, it is convenient to abbreviate their norms. Thus let

$$\|\cdot\|_s=\|\cdot\|_{H^s(\mathbb{R}^+)} \quad \text{and} \quad |\cdot|_{s,T}=\|\cdot\|_{H^s(0,T)}. \tag{2.2a}$$

If $s=0$, the subscript s will be omitted altogether. So

$$\|\cdot\|=\|\cdot\|_{L^2(\mathbb{R}^+)} \quad \text{and} \quad |\cdot|_T=|\cdot|_{0,T}. \tag{2.2b}$$

Some special cases of the Sobolev embedding theorems will be used occasionally and are worth recalling here. Let I be an open interval on the real line, not necessarily bounded. If $s>1/2+m$, where m is a nonnegative integer, then

$$H^s(I)\subset C_b^m(\bar{I}), \tag{2.3}$$

algebraically, and continuously with respect to the norms on these two spaces. (More precisely, an element in $H^s(I)$ is, after possible modification on a set of Lebesgue measure zero, a C^m-function on I, all of whose derivatives up to order m are uniformly continuous on I, and so may be extended to $\bar{I}$.) In the special case where $I=\mathbb{R}^+$ and $s=k$, a positive integer, it is also useful to recall that if $f\in H^k(\mathbb{R}^+)$, then,

$$f(x),f'(x),\cdots,f^{(k-1)}(x)\to 0 \quad \text{as } x\to+\infty. \tag{2.4}$$

An inequality that will find use is the following, valid for $f\in H^1(\mathbb{R}^+)$. According to (2.3), such a function is bounded and continuous on $\mathbb{R}^+$, and furthermore,

$$\|f\|_{C_b(\mathbb{R}^+)}\le\sqrt{2}\,(\|f\|\|f'\|)^{1/2}. \tag{2.5}$$

This inequality, which is sharp in fact, follows from the observation that, for any $y\in\mathbb{R}^+$, and $f\in H^1(\mathbb{R}^+)$,

$$f^2(y)=-2\int_y^\infty f(x)f'(x)\,dx\le 2\left\{\int_y^\infty f^2(x)\,dx\cdot\int_y^\infty [f'(x)]^2dx\right\}^{1/2}$$
$$\le 2\|f\|\|f'\|.$$

Spaces will be needed to describe the evolution in time of the spatial structure. If X is a Banach space, $1\le p\le\infty$, and $-\infty\le a<b\le\infty$, then $L^p(a,b;X)$ denotes the space of measurable functions $u:(a,b)\to X$ whose norms are pth-power integrable (essentially bounded, if $p=\infty$). These are Banach spaces in their own right, with the norms

$$\|u\|_{L^p(a,b;X)}=\left\{\int_a^b\|u(t)\|_X^p\,dt\right\}^{1/p} \quad \text{for } p<\infty,$$

and

$$\|u\|_{L^\infty(a,b;X)}=\operatorname*{essential\,supremum}_{t\in(a,b)}\{\|u(t)\|_X\}.$$

The subspace of $L^\infty(a,b;X)$ of continuous and bounded functions $u:[a,b]\to X$ is denoted $C_b(a,b;X)$. (In case a and b are both finite, the subscript b, for "bounded", is dropped.)

These spaces all possess localized versions. The only one appearing here is the space $L^\infty_{\text{loc}}(\overline{\mathbb{R}}^+;X)$ of measurable maps $u:\overline{\mathbb{R}}^+\to X$ which are essentially bounded on any compact subset of $\overline{\mathbb{R}}^+$.

Finally, if X is still an arbitrary Banach space, we may consider the X-valued distributions $\mathcal{D}'(a,b;X)$ on the interval (a,b). Formally, $\mathcal{D}'(a,b;X)$ is the set of linear

and continuous maps of $\mathcal{D}(a,b)$ into X. If $T\in\mathcal{D}'(a,b;X)$, its distributional derivative is defined by

$$\frac{dT}{dt}(\varphi)=-T(\varphi'),$$

for $\varphi\in\mathcal{D}(a,b)$. Thus, if $f\in L^p(a,b;X)$, then f may be viewed as an X-valued distribution via the definition

$$f(\varphi)=\int_a^b f(t)\varphi(t)\,dt,$$

for $\varphi\in\mathcal{D}(a,b)$. The integral is, of course, X-valued, and converges since φ has compact support. Thus, "temporal" derivatives of $L^p(a,b;X)$-functions may always be defined, at least in the distributional sense. There is a considerable theory pertaining to when distributional derivatives are in fact classically defined. Some of these results will be called upon later. Specific uses of this theory will be referenced precisely, but the reader may consult [18], [19], [25] or [28] for general commentary concerning such issues.

The following is a special case of the main result of this paper. It serves simultaneously to give orientation and define the goals of the paper.

THEOREM. *Consider the initial- and boundary-value problem* (1.3) *and suppose that the data* f,g *has* $f\in H^4(\mathbb{R}^+)$ *and* $g\in H^2_{\text{loc}}(\mathbb{R}^+)$. *Suppose that* f *and* g *satisfy the compatibility conditions,*

$$g(0)=f(0),$$
$$g_t(0)=-\left(f_{xxx}(0)+f(0)f_x(0)+f_x(0)\right).$$

Then there exists a unique solution u *in* $L^\infty_{\text{loc}}(\mathbb{R}^+;H^4(\mathbb{R}^+))$ *of* (1.3) *corresponding to the data* f *and* g.

Remarks. By the term "solution", we will always mean, in the first instance, a solution in the sense of distributions on the quarter-plane. The term *classical solution* is reserved for a function which is continuous and continuously differentiable the requisite number of times, and which satisfies the differential equation pointwise everywhere, and the initial and the boundary conditions pointwise.

Note that since $g\in H^2_{\text{loc}}(\mathbb{R}^+)$, $g\in C^1(0,T)$, for any $T>0$. Also, $f\in H^4(\mathbb{R}^+)$ implies $f\in C^3_b(\overline{\mathbb{R}}^+)$. In consequence, the compatibility conditions are both well-defined. The first compatibility condition simply expresses the continuity of the solution u at the origin. The second condition would necessarily hold for a classical solution.

The theorem above is a part of Theorem 6.2 below. There it will also be established that if $f\in H^{3k+1}(\mathbb{R}^+)$ and $g\in H^{k+1}_{\text{loc}}(\mathbb{R}^+)$, where k is a positive integer, and if corresponding higher order compatibility conditions hold, then the solution u lies in the class $L^\infty_{\text{loc}}(\mathbb{R}^+;H^{3k+1}(\mathbb{R}^+))$. In particular, if $k\geq 2$, it is easily inferred that u is a classical and global solution of the quarter-plane problem for the KdV equation.

3. Theory relating to the regularized problem. In this section, interest will be focused entirely on the regularized initial- and boundary-value problem (1.5), repeated here for convenience.

$$u_t+u_x+uu_x+u_{xxx}-\varepsilon u_{xxt}=0 \quad \text{for } x,t\geq 0, \tag{3.1a}$$

with

$$u(x,0)=f(x) \quad \text{for } x\geq 0, \tag{3.1b}$$
$$u(0,t)=g(t) \quad \text{for } t\geq 0.$$

For consistency, the restriction

$$u(0,0)=f(0)=g(0) \tag{3.2}$$

will be imposed throughout the discussion. For the present, the positive parameter ε will be treated as a fixed constant, in the range $(0,1]$, say. Following the development in [8], let

$$v(x,t)=\varepsilon u\left(\varepsilon^{1/2}(x-t),\varepsilon^{3/2}t\right). \tag{3.3}$$

It is immediately verified that u is a smooth solution of (3.1) if and only if v is a smooth solution of the problem

$$v_t+(1+\varepsilon)v_x+vv_x-v_{xxt}=0 \quad \text{in } \bar{\Omega}, \tag{3.4a}$$

and

$$\begin{aligned} &v(x,0)=F(x) \quad \text{for } x\geq 0, \\ &v(t,t)=G(t) \quad \text{for } t\geq 0. \end{aligned} \tag{3.4b}$$

Here $\Omega=\{(x,t)\colon t>0 \text{ and } x>t\}$, $F(x)=\varepsilon f(\varepsilon^{1/2}x)$, and $G(t)=\varepsilon g(\varepsilon^{3/2}t)$. The dependence of F and G on ε is suppressed, since ε is viewed as fixed here. Of course (3.2) implies and is implied by

$$F(0)=G(0). \tag{3.5}$$

The initial- and boundary-value problem (3.4) is somewhat peculiar, owing to the domain (a sector of angle $\pi/4$) in which it is posed (cf. Fig. 2).

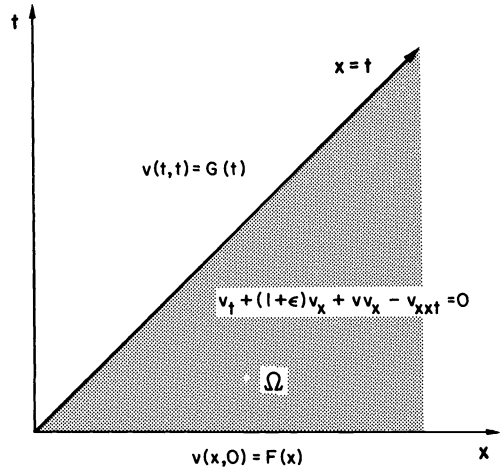


FIG. 2. *The regularized problem, after the change of variables.*

Related initial- and boundary-value problems have been analyzed by passing to an associated integral equation. This method proves to be effective in the present circumstances.

To convert (3.4) into an integral equation, proceed formally as follows. Write (3.4) as

$$v_t-v_{xxt}=-(1+\varepsilon)v_x-vv_x,$$

and, for fixed $x \geq t$, integrate this relation over the temporal interval $(0,t)$. There appears

$$w - w_{xx} = S \quad \text{for } x > t, \tag{3.6}$$

where

$$w(x,t) = v(x,t) - F(x),$$

and

$$S(x,t) = -\int_0^t [(1+\varepsilon)v_x(x,s) + v(x,s)v_x(x,s)]\, ds.$$

The solution of (3.6) may be expressed in the form

$$w(x,t) = \alpha e^{-x} + \frac{1}{2}\int_t^\infty e^{-|x-\xi|} S(\xi,t)\, d\xi, \tag{3.7}$$

by the variation of constants formula. Of course $\alpha = \alpha(t)$, and it has been assumed tacitly that S and w are bounded. If $t \geq 0$, then at $x = t$,

$$G(t) - F(t) = v(t,t) - F(t) = w(t,t) = \alpha(t)e^{-t} + \frac{1}{2}\int_t^\infty e^{-|t-\xi|} S(\xi,t)\, d\xi.$$

Hence,

$$\alpha(t) = e^t \left\{ G(t) - F(t) - \frac{1}{2}\int_t^\infty e^{-|t-\xi|} S(\xi,t)\, d\xi \right\}. \tag{3.8}$$

The result of (3.7) and (3.8) is that

$$v(x,t) = F(x) + e^{-(x-t)}(G(t) - F(t)) - \frac{1}{2}e^{-(x-t)}\int_t^\infty e^{-|t-\xi|} S(\xi,t)\, d\xi + \frac{1}{2}\int_t^\infty e^{-|x-\xi|} S(\xi,t)\, d\xi.$$

Since $\xi \geq t$, this simplifies to

$$v(x,t) = F(x) + e^{-(x-t)}(G(t) - F(t)) + \int_t^\infty M(x-t, \xi-t) S(\xi,t)\, d\xi, \tag{3.9}$$

where

$$M(y,z) = \frac{1}{2}[\exp(-|y-z|) - \exp(-(y+z))]. \tag{3.10}$$

Replacing S by its definition in terms of v, and integrating once by parts, (3.9) may be expressed in the form

$$v(x,t) = F(x) + e^{-(x-t)}(G(t) - F(t)) + \int_t^\infty K(x-t, \xi-t)\int_0^t \left[(1+\varepsilon)v(\xi,s) + \frac{1}{2}v^2(\xi,s)\right] ds\, d\xi, \tag{3.11}$$

where

$$K(y,z) = \frac{1}{2}[\exp(-y-z) + \operatorname{sgn}(y-z)\exp(-|y-z|)]. \tag{3.12}$$

The boundary term that appears in the integration by parts vanishes because $e^{-|x-\xi|} = e^{-(x+\xi)+2t}$ when $\xi = t$ and $x \geq t$. Notice that $K(0, \xi - t) \equiv 0$, so that $v(t,t) = G(t)$, for all $t \geq 0$. Note also that $v(x,0) = F(x)$, provided the consistency condition (3.5) holds.

Equation (3.11) is the desired integral equation. It has been derived formally, and thus far its relation to solutions of (3.4) is not rigorously established. Our object now is to make a rigorous connection between solutions of the integral equation and solutions of (3.4), and to show that the integral equations possesses solutions, at least for small time intervals.

Turning to the second objective first, let $T>0$ and let $\mathcal{C}_T$ be the Banach space of bounded continuous functions defined on the closure of the set

$$\Omega_T=\{(x,t)\colon t\in(0,T) \text{ and } x>t\}.$$

$\mathcal{C}_T$ is equipped with the supremum norm. Let A denote the operator that maps a function $w\in\mathcal{C}_T$ into the function

$$(3.13)\quad (Aw)(x,t)=F(x)+e^{-(x-t)}(G(t)-F(t)) + \int_t^\infty K(x-t,\xi-t)\int_0^t\left[(1+\varepsilon)w(\xi,s)+\frac{1}{2}w^2(\xi,s)\right]ds\,d\xi,$$

defined for $(x,t)\in\overline{\Omega}_T$. Because the kernel K is integrable, and assuming that F and G are bounded and continuous, it is plain that $Aw\in\mathcal{C}_T$ also. Existence of a solution of the integral equation (3.11) will be provided by showing that, for T small enough, A is a contraction mapping of a ball centered at the zero function in $\mathcal{C}_T$. The following estimate is the basis on which this assertion is established.

Let u and w be elements of $\mathcal{C}_T$. Consider the difference of their images under the operator A,

$$Au(x,t)-Aw(x,t) =\int_t^\infty K(x-t,\xi-t)\int_0^t\left[1+\varepsilon+\frac{1}{2}u(\xi,s)+\frac{1}{2}w(\xi,s)\right][u(\xi,s)-w(\xi,s)]\,ds\,d\xi.$$

For t fixed in the interval $[0,T]$,

$$\sup_{x\geq t}|Au(x,t)-Aw(x,t)| \leq\sup_{x\geq t}\int_t^\infty|K(x-t,\xi-t)|\,d\xi\sup_{\xi\geq t}\int_0^t\left|1+\varepsilon+\frac{1}{2}u(\xi,s)+\frac{1}{2}w(\xi,s)\right||u(\xi,s)-w(\xi,s)|\,ds.$$

But, for $x\geq t$,

$$\int_t^\infty|K(x-t,\xi-t)|\,d\xi=\frac{1}{2}\int_t^\infty|e^{2t-(x+\xi)}+\operatorname{sgn}(x-\xi)e^{-|x-\xi|}|\,d\xi$$
$$=\frac{1}{2}\int_x^\infty|e^{2t-(x+\xi)}-e^{x-\xi}|\,d\xi+\frac{1}{2}\int_t^x(e^{2t-(x+\xi)}+e^{\xi-x})\,d\xi$$
$$=1-e^{2(t-x)}\leq 1.$$

Hence, as $0\leq t\leq T$,

$$\sup_{x\geq t}|Au(x,t)-Aw(x,t)|\leq\sup_{\xi\geq t}\int_0^t\left|1+\varepsilon+\frac{1}{2}u(\xi,s)+\frac{1}{2}w(\xi,s)\right||u(\xi,s)-w(\xi,s)|\,ds$$
$$\leq T\left[1+\varepsilon+\frac{1}{2}(\|u\|_{\mathcal{C}_T}+\|w\|_{\mathcal{C}_T})\right]\|u-w\|_{\mathcal{C}_T}.$$

It follows that

$$\|Au-Aw\|_{\mathcal{C}_T}=\sup_{(x,t)\in\Omega_T}|Au(x,t)-Aw(x,t)| \tag{3.14}$$
$$\leq T\left[(1+\varepsilon)+\frac{1}{2}\left(\|u\|_{\mathcal{C}_T}+\|w\|_{\mathcal{C}_T}\right)\right]\|u-w\|_{\mathcal{C}_T}.$$

This inequality implies the desired result. Let $\theta(x,t)\equiv 0$ and set

$$R(T)=2\|A\theta\|_{\mathcal{C}_T}\leq 4\|F\|_{C_b(\bar{\mathbb{R}}^+)}+2\|G\|_{C(0,T)}. \tag{3.15}$$

Let $B_T=\{w\in\mathcal{C}_T\colon \|w\|_{\mathcal{C}_T}\leq R(T)\}$ and let

$$\Theta(T)=T[1+\varepsilon+R(T)]. \tag{3.16}$$

Then it follows straightforwardly that, for u and w in B_T,

$$\|Au-Aw\|_{\mathcal{C}_T}\leq\Theta(T)\|u-w\|_{\mathcal{C}_T},$$

and

$$\|Au\|_{\mathcal{C}_T}\leq\|Au-A\theta\|_{\mathcal{C}_T}+\|A\theta\|_{\mathcal{C}_T}\leq\Theta(T)\|u\|_{\mathcal{C}_T}+\frac{1}{2}R(T)\leq\left[\Theta(T)+\frac{1}{2}\right]R(T).$$

Because of the last two inequalities, A will be a contraction mapping of B_T if $\Theta(T)\leq\frac{1}{2}$. Referring to (3.16), one appreciates immediately that, for fixed data F and G, this certainly holds for T sufficiently small. In fact, it is worth noting that, essentially because of the inequality in (3.15), for any $M>0$ we may take

$$T=\min\left\{M,\frac{1}{2\left(1+\varepsilon+4\|F\|_{C_b(\bar{\mathbb{R}}^+)}+2\|G\|_{C(0,M)}\right)}\right\}, \tag{3.17}$$

and have $\Theta(T)\leq\frac{1}{2}$. Thus (3.11) has a solution in $\mathcal{C}_T$, for T sufficiently small. This result is summarized formally in the following.

PROPOSITION 3.1. *Let $M>0$, $G\in C(0,M)$ and $F\in C_b(\bar{\mathbb{R}}^+)$ with $F(0)=G(0)$. Then there exists a positive constant*

$$T_0=T_0\left(\|F\|_{C_b(\bar{\mathbb{R}}^+)},\|G\|_{C(0,M)}\right)$$

such that for any T' with $0<T'\leq\min(T_0,M)$, there is a solution of (3.11) *in $\mathcal{C}_{T'}$. Moreover, for any $T\in(0,M]$, there is at most one solution of* (3.11) *in $\mathcal{C}_T$.*

Proof. The question of existence has already been settled. Suppose there are two distinct solutions v and w of (3.11) in $\mathcal{C}_T$. Since v and w are continuous, there is a $t_0\in[0,T)$ such that $v\equiv w$ on Ω_{t_0}, and on no domain Ω_t is this still true, if $t>t_0$. Let $U(x,t)=v(x,t)=w(x,t)$, in $\bar{\Omega}_{t_0}$. Define

$$U_0(x,t)=F(x)+e^{-(x-t)}(G(t)-F(t))$$
$$+\int_t^\infty K(x-t,\xi-t)\int_0^{t_0}\left[(1+\varepsilon)U(\xi,s)+\frac{1}{2}U^2(\xi,s)\right]ds\,d\xi,$$

for $(x,t)\in D=\{(x,t)\colon t_0\leq t\leq T \text{ and } x\geq t\}$. Plainly U_0 is bounded and continuous on D. Then the integral equation

$$u(x,t)=U_0(x,t)+\int_t^\infty K(x-t,\xi-t)\int_{t_0}^t\left[(1+\varepsilon)u(\xi,s)+\frac{1}{2}u^2(\xi,s)\right]ds\,d\xi$$
$$=\tilde{A}u(x,t),$$

defined on D, has two distinct solutions, which we denote by v and w again, though they are in fact v and w restricted to D. Moreover, while these two solutions agree at t_0, they do not agree identically in any neighborhood of t_0.

The existence argument presented above is easily adapted to show that, for R large enough and for $t_1 = t_1(R)$ near enough to t_0, $\tilde{A}$ is a contraction mapping of the ball $\tilde{B}_R$ of radius R centered at the zero function in $C_b(D_1)$, where

$$D_1 = \{(x,t)\colon t_0 \le t \le t_1 \text{ and } x \ge t\}.$$

But if

$$R \ge \max\{\|v\|_{\mathcal{C}_T}, \|w\|_{\mathcal{C}_T}\},$$

then $\tilde{A}$ has two distinct fixed points v and w in $\tilde{B}_R$. This contradiction forces the conclusion $v \equiv w$ on Ω_T, and the proposition is established. □

It will be important in subsequent sections to have smooth solutions, up to the boundaries, of the regularized problem (3.1) at our disposal. This amounts to the program of relating solutions of the integral equation (3.11) to solutions of the transformed problem (3.4). The following result will be sufficient for our later needs.

PROPOSITION 3.2. *Suppose that* $F \in C_b^k(\overline{\mathbb{R}}^+)$ *and* $G \in C^m(0, T_0)$, *where* $k \ge 2$, $m \ge 1$, *and* $k \ge m$. *Suppose also* $F(0) = G(0)$. *Let* v *be a solution in* $\mathcal{C}_T$ *of the integral equation* (3.11), *where* $0 < T \le T_0$. *Then*

$$\partial_x^i \partial_t^j v \in \mathcal{C}_T, \quad \text{for } 0 \le j \le m \quad \text{and} \quad 0 \le i \le k + j. \tag{3.18}$$

Moreover, v *is a classical solution of the transformed problem* (3.4) *in* $\overline{\Omega}_T$. *Conversely, if* v *lies in* $\mathcal{C}_T$ *and is a classical solution of* (3.4) *on* $\overline{\Omega}_T$, *then* v *is a solution of the integral equation* (3.11) *over* $\overline{\Omega}_T$, *and so* v *satisfies* (3.18).

Remark. The partial derivatives in (3.18) may be defined at the boundary of Ω_T by the obvious one-sided differential quotients. The reader will appreciate that a function v defined on $\overline{\Omega}_T$ does not possess a classically defined partial derivative with respect to t at the point $(0,0)$. In case $j > 0$ in (3.18), the condition $\partial_x^i \partial_t^j v \in \mathcal{C}_T$ connotes that this partial derivative exists classically in $\overline{\Omega}_T \setminus \{(0,0)\}$, is bounded and continuous there, and that it may be extended continuously to $\overline{\Omega}_T$.

Proof. First note that if $F \in C_b^k(\overline{\mathbb{R}}^+)$ and $G \in C^m(0,T)$, where $k \ge m$, then

$$v_0(x,t) = F(x) + e^{-x}e^t(G(t) - F(t)) \tag{3.19}$$

has $\partial_x^i \partial_t^j v_0 \in \mathcal{C}_T$, for $0 \le i \le k$ and $0 \le j \le m$. Also, since $v \in \mathcal{C}_T$, then

$$J(x,t) = \int_0^t \left[(1+\varepsilon)v(x,s) + \frac{1}{2}v^2(x,s)\right] ds \tag{3.20}$$

has $J_t \in \mathcal{C}_T$. A short calculation using Leibniz' rule confirms that

$$\begin{aligned} v_t(x,t) = {} & \partial_t v_0(x,t) - K(x-t,0)J(t,t) \\ & + \int_t^\infty \partial_t[K(x-t,\xi-t)]J(\xi,t)\,d\xi + \int_t^\infty K(x-t,\xi-t)J_t(\xi,t)\,d\xi. \end{aligned}$$

Simplifying,

$$\begin{aligned} v_t(x,t) = {} & \partial_t v_0(x,t) - e^{-(x-t)}J(t,t) \\ & + \int_t^\infty e^{2t-(x+\xi)}J(\xi,t)\,d\xi + \int_t^\infty K(x-t,\xi-t)J_t(\xi,t)\,d\xi. \end{aligned} \tag{3.21}$$

Thus $v_t \in \mathcal{C}_T$.

By dividing the range of spatial integration at $\xi=x$, it is readily seen that $v_x\in\mathcal{C}_T$, and that

$$v_x(x,t)=\partial_x v_0(x,t)+K_-(x-t,x-t)J(x,t)-K_+(x-t,x-t)J(x,t)+\int_t^\infty L(x-t,\xi-t)J(\xi,t)\,d\xi, \tag{3.22}$$

where

$$L(y,z)=-\frac{1}{2}\{\exp(-|y-z|)+\exp(-y-z)\}, \tag{3.23}$$

$$K_\pm(x-t,x-t)=\lim_{\xi\to x\pm}K(x-t,\xi-t),$$

and $\xi\to x+$ means $\xi\downarrow x$ while $\xi\to x-$ means $\xi\uparrow x$. Thus it appears that

$$v_x(x,t)=\partial_x v_0(x,t)+J(x,t)+\int_t^\infty L(x-t,\xi-t)J(\xi,t)\,d\xi. \tag{3.24}$$

Since $k\geq 2$, $\partial_x v_0$ may be differentiated with respect to x. Moreover, since $v_x\in\mathcal{C}_T$, $J(x,t)$ may be differentiated with respect to x. And, the integral on the right side of (3.24) may be differentiated with respect to x. Performing the indicated differentiations, we see that

$$v_{xx}(x,t)=\partial_x^2 v_0(x,t)+J_x(x,t)+\int_t^\infty K(x-t,\xi-t)J(\xi,t)\,d\xi. \tag{3.25}$$

This representation shows plainly that $v_{xx}\in\mathcal{C}_T$. Formula (3.25) may be simplified by use of the original integral equation. Thus,

$$\begin{aligned} v_{xx}(x,t)&=\partial_x^2 v_0(x,t)+J_x(x,t)+(v(x,t)-v_0(x,t))\\ &=J_x(x,t)+v(x,t)+F_{xx}(x)-F(x)\\ &=\int_0^t[(1+\varepsilon)v_x(x,s)+v(x,s)v_x(x,s)]\,ds+v(x,t)+F_{xx}(x)-F(x). \end{aligned} \tag{3.26}$$

It is now clear that v_{xx} is differentiable with respect to t, and that

$$v_{xxt}(x,t)=(1+\varepsilon)v_x(x,t)+v(x,t)v_x(x,t)+v_t(x,t).$$

So, if $k\geq 2$ and $m\geq 1$, any solution v in $\mathcal{C}_T$ of the integral equation (3.11) is a classical solution, up to the boundary, of the transformed differential equation (3.4a). As already remarked, a continuous solution of (3.11) has $v(t,t)=G(t)$, for $0\leq t\leq T$, and has $v(x,0)=F(x)$, for $x\geq 0$, provided the consistency condition $F(0)=G(0)$ holds.

Further regularity of a $\mathcal{C}_T$-solution of the integral equation may be established by similar arguments. As this issue is important in our subsequent investigation, a little more detail is warranted.

First, if $m\geq 2$, then since $v_t\in\mathcal{C}_T$, it follows that every term on the right-hand side of (3.21) is differentiable with respect to t. Moreover, each of these derivatives lies in $\mathcal{C}_T$, as is easily verified. So $v_{tt}\in\mathcal{C}_T$. This argument may now be iterated, with the conclusion that $\partial_t^j v\in\mathcal{C}_T$, for $0\leq j\leq m$.

A similar argument, based on (3.26), may be used to show that $\partial_x^i v\in\mathcal{C}_T$, for $0\leq i\leq k$. Specifically,

$$\partial_x^{l+2}v(x,t)=\partial_x^l v(x,t)+\partial_x^{l+2}F(x)-\partial_x^l F(x)+\int_0^t\partial_x^{l+1}\left[(1+\varepsilon)v(x,s)+\frac{1}{2}v^2(x,s)\right]ds, \tag{3.27}$$

for $l=0,1,\cdots,k-2$.

Since $v_t \in \mathcal{C}_T$, it follows from (3.24) that $v_{xt} \in \mathcal{C}_T$ and that

$$(3.28)\quad v_{xt}(x,t) = \partial_t \partial_x v_0(x,t) + J_t(x,t) + e^{-(x-t)}J(t,t) + \int_t^\infty L(x-t,\xi-t)J_t(\xi,t)\,d\xi - \int_t^\infty e^{2t-(x+\xi)}J(\xi,t)\,d\xi.$$

Finally, by using the differential equation, the results already derived, and induction, mixed partial derivatives of the form $\partial_x^i \partial_t^j v$, where $j \geq 1$ and $i \geq 2$, are seen to lie in $\mathcal{C}_T$, provided that $j \leq m$ and $i \leq k+j$.

If, on the other hand, v is a bounded classical solution of the differential equation (3.4a) which satisfies the boundary conditions (3.4b), then necessarily $F(0) = G(0)$ because v is continuous at the origin. Moreover, in this case, each step of the formal construction leading from (3.1) to (3.4) is easily validated. In consequence, v is seen to satisfy (3.11). Hence by the argument just elucidated, pertaining to solutions of the integral equation (3.11), v satisfies the conditions of regularity in (3.18). This concludes the proof of the proposition. □

In our subsequent analysis, it will be convenient to have at our disposal smooth solutions of (3.1) which are not confined to $\overline{\mathbb{R}}^+ \times [0,T]$ where T is small. This corresponds to providing smooth solutions of (3.4) on $\overline{\Omega}_{T'}$, where T' is given. It seems natural to iterate the local result propounded in Proposition 1. This will be effective as soon as an a priori bound on the L^∞-norm of a solution defined on $\overline{\Omega}_T$ is provided. More precisely, suppose a classical solution v of (3.4), defined on $\overline{\Omega}_T$ for some $T > 0$, is in hand. And suppose the boundary data G is defined at least on $[0, T_0]$, where $T_0 > T$. Consider a new initial- and boundary-value problem,

$$(3.29)\quad w_t + (1+\varepsilon)w_x + ww_x - w_{xxt} = 0 \quad \text{for } (x,t) \text{ such that } t \geq T \text{ and } x \geq t,$$

with

$$w(x,T) = v(x,T) \quad \text{for } x \geq T,$$
$$w(t,t) = G(t) \quad \text{for } t \geq T.$$

The initial value of w is the terminal value of v. Just as for (3.4), (3.29) may be converted to an integral equation, which in all aspects is similar to (3.11). A solution to this integral equation may be inferred to exist on some domain of the form

$$\{(x,t)\colon T \leq t \leq T + \Delta T \text{ and } x \geq t\}.$$

Provided v and G are smooth enough, the solution w of the integral equation will provide a classical solution of (3.29). In this manner, v is extended to a solution of (3.4) on $\overline{\Omega}_{T+\Delta T}$. As in Proposition 1, a lower bound for the size of ΔT depends on the L^∞-norm of the data in (3.29). Specifically referring to (3.17),

$$\Delta T \geq \min\left\{ T_0 - T,\ \frac{1}{2\left[1 + \varepsilon + 4\|v(\cdot,T)\|_{C_b(T,\infty)} + 2\|G\|_{C(T,T_0)}\right]} \right\}.$$

Suppose it is known that, for the given data F and G, any solution v of (3.4) defined on $\overline{\Omega}_T$, for some $T \leq T_0$, has the property that

$$\|v\|_{C_b(\overline{\Omega}_T)} \leq C = C(T_0, F, G).$$

Then a lower bound on ΔT can be imputed, and in consequence, after a finite number of steps, the solution may be extended to $\overline{\Omega}_{T_0}$. This conclusion is worth stating formally.

PROPOSITION 3.3. *Let $T_0>0$ be given, and $G\in C^m(0,T_0)$, $F\in C_b^k(\overline{\mathbb{R}}^+)$ with $F(0)=G(0)$, where $k\geq 2$, $m\geq 1$ and $k\geq m$. Suppose there is a constant C, dependent on T_0, F and G, such that for any solution w of* (3.4) *defined on $\overline{\Omega}_T$, where $T\leq T_0$,*

$$\|w\|_{C_b(\overline{\Omega}_T)}\leq C. \tag{3.30}$$

Then there exists a unique solution $v\in\mathcal{C}_{T_0}$ to (3.11), *which is also a classical solution of* (3.4) *and which satisfies the conditions of regularity expressed in* (3.18). *Moreover, v is defined locally as the fixed-point of a contraction mapping of the type in* (3.13), *by iterating the result of Proposition* 3.1 *a finite number of times.*

Provision of the relevant a priori bound is now considered. To this end, the following technical lemma is useful.

LEMMA 3.4. *Let $F\in C_b^k(\overline{\mathbb{R}}^+)$ and $G\in C^m(0,T_0)$ with $F(0)=G(0)$, where $k\geq 2$, $m\geq 1$ and $k\geq m$. Let v be a solution of* (3.4) *in $\mathcal{C}_{T_0}$. Let $0\leq p\leq k$ and suppose that*

$$\partial_x^j F(x)\to 0 \quad \text{as } x\to+\infty, \tag{3.31}$$

for $0\leq j\leq p$. Then

$$\partial_x^j\partial_t^i v(x,t)\to 0 \quad \text{as } x\to+\infty, \tag{3.32}$$

uniformly for $0\leq t\leq T_0$, for i,j such that $0\leq i\leq m$ and $0\leq j\leq p+i$.

Proof. Suppose it is determined that $v(x,t)\to 0$ as $x\to+\infty$, uniformly for $0\leq t\leq T_0$. Since v is a classical solution of (3.4) on $\overline{\Omega}_{T_0}$, it satisfies the integral equation (3.11) on $\overline{\Omega}_{T_0}$. Referring to formula (3.21) for v_t, it is clear that $v_t(x,t)\to 0$ as $x\to+\infty$, uniformly for $0\leq t\leq T_0$. If $m>1$, then upon differentiating (3.21) with respect to t and using the fact that v and v_t tend to 0 at $+\infty$, it is straightforwardly assured that $v_{tt}(x,t)\to 0$ as $x\to\infty$, uniformly for $0\leq t\leq T_0$. Continuing inductively, it follows that

$$\partial_t^i v(x,t)\to 0 \quad \text{as } x\to+\infty,$$

for $0\leq i\leq m$, uniformly for $0\leq t\leq T_0$.

Next, by considering formula (3.22), we see that if $p>0$, then $v_x(x,t)\to 0$ as $x\to+\infty$, uniformly for $0\leq t\leq T_0$. Then from (3.28), $v_{xt}(x,t)\to 0$ as $x\to+\infty$, uniformly for $0\leq t\leq T_0$. From the differential equation (3.4a), it is seen that $v_{xxt}(x,t)\to 0$, as $x\to+\infty$, uniformly for $0\leq t\leq T_0$. Continuing in the pattern of the proof of Proposition 3.2 leads to the conclusion that (3.32) holds.

The above analysis was all predicated on the desired result holding good for v itself. The lemma will therefore be established as soon as it is confirmed that (3.32) holds for $i=j=0$.

For $T>0$, let $\mathcal{C}_T^0$ be the closed subspace of $\mathcal{C}_T$ composed of those elements which converge to 0 at $+\infty$, uniformly for $0\leq t\leq T$. If $F(x)\to 0$, as $x\to+\infty$, then operators of the type exhibited in (3.13) map $\mathcal{C}_T^0$ into itself. Because a solution v of (3.4) is provided in $\mathcal{C}_{T_0}$, the uniqueness result of Proposition 3.1 implies that condition (3.30) holds. So v is obtained locally as a fixed-point of a contraction mapping of the form in (3.13). This fixed-point may be determined by iterating the operator on the zero function θ. The sequence $\{v_n\}_{n=1}^{\infty}$ thus generated ($v_1=A\theta$ and $v_{n+1}=Av_n$, for $n\geq 1$) lies in $\mathcal{C}_T^0$ and converges to v in $\mathcal{C}_T$. Therefore $v\in\mathcal{C}_T^0$. As a finite number of such steps are needed to recover v on $\overline{\Omega}_{T_0}$, it follows that $v\in\mathcal{C}_{T_0}^0$. This concludes the proof of the lemma. □

Attention is now turned fully toward derivation of a priori information concerning smooth solutions of (3.4) which imply (3.30). A bound that will suffice is the subject of the next proposition. The same bound will also be needed in §4. Because of this, it is especially convenient to derive the bound in the context of (3.1). Of course the reader

will realize that the theory, thus far developed for (3.4), implies the existence of smooth solutions of the regularized problem (3.1), at least locally in time. This is simply a matter of tracing the inverse of the transformation (3.3) which led from (3.1) to (3.4). The precise result is spelled out in Theorem 3.8. For now, it is simply assumed that a classical solution of (3.1) is in hand.

PROPOSITION 3.5. *Let* $f\in C_b^3(\bar{\mathbf{R}}^+)$, $g\in C^1(0,T)$, *where* $f(0)=g(0)$, *and suppose* $0<\varepsilon\le 1$. *Let* u *be a classical solution of* (3.1), *up to the boundary, on* $\bar{\mathbf{R}}^+\times[0,T]$. *Suppose in addition that* $f\in H^1(\mathbf{R}^+)$. *Then for all* $t\in[0,T]$, $u(\cdot,t)\in H^1(\mathbf{R}^+)$. *Moreover, there are positive constants* a_0 *and* a_1,

$$a_0=a_0\big(\|f\|+\varepsilon^{1/2}\|f_x\|,|g|_{1,T}\big)$$

and

$$a_1=a_1\big(\|f\|_1,|g|_{1,T}\big),$$

depending continuously on their arguments, such that

$$\|u(\cdot,t)\|\le a_0 \tag{3.33}$$

and

$$\|u(\cdot,t)\|_1^2+\int_0^t\big[u_x^2(0,s)+\big(u_{xx}(0,s)-\varepsilon u_{xt}(0,s)\big)^2\big]\,ds\le a_1, \tag{3.34}$$

for $0\le t\le T$. *These inequalities hold uniformly for* ε *in* $(0,1]$.

Remark. While not stated explictly here or later, the various constants that appear in the development of our theory generally depend on T. Besides a direct dependence on T, a_0 and a_1 also depend indirectly on T via the $H^1(0,T)$-norm of g, $|g|_{1,T}$. The reader will quickly perceive that a_0 and a_1 may be presumed to depend monotonically on T, for given f and g. In fact, a_0 and a_1 may be assumed to depend monotonically on their arguments generally, but this will not be needed here.

Before proving the proposition, the following corollary result is stated. This is the result of central interest for the present section.

COROLLARY 3.6. *Let* $F\in C_b^3(\bar{\mathbf{R}}^+)$ *and* $G\in C^1(0,T_0)$ *with* $F(0)=G(0)$. *Suppose in addition that* $F\in H^1(\mathbf{R}^+)$. *Then there exists a constant* C, *dependent on* $\|F\|_1$ *and the* $H^1(0,T_0)$*-norm of* G, *such that any classical solution* v *of* (3.4) *defined on* $\bar{\Omega}_T$, *for* $T\le T_0$, *satisfies*

$$\|v\|_{C_b(\bar{\Omega}_T)}\le C.$$

Proof. Let v be a classical solution of (3.4) on $\bar{\Omega}_T$, for some $T\le T_0$. The inverse of the change of variables (3.3) is

$$u(x,t)=\varepsilon^{-1}v(\varepsilon^{-1/2}x+\varepsilon^{-3/2}t,\varepsilon^{-3/2}t). \tag{3.35}$$

Then u is a classical solution of (3.1a) on $\bar{\mathbf{R}}^+\times[0,T']$, where $T'=\varepsilon^{-3/2}T$, which satisfies the auxiliary conditions (3.16) where

$$f(x)=\varepsilon^{-1}F(\varepsilon^{-1/2}x)\quad\text{and}\quad g(t)=\varepsilon^{-1}G(\varepsilon^{-3/2}t). \tag{3.36}$$

Here $\varepsilon>0$ is fixed, and so f and g satisfy the hypotheses of Proposition 3.5. Hence the $H^1(\mathbf{R}^+)$-norm of u is bounded on $[0,T']$ by a constant that depends on $\|f\|_1$, and on the $H^1(0,T_0')$-norm $|g|_{1,T_0'}$ of g, say. Here, $T_0'=\varepsilon^{3/2}T_0$. Because of the basic inequality (2.1), it follows that u is bounded on $\bar{\mathbf{R}}^+\times[0,T']$ by a constant C dependent only on $\|f\|_1$ and $|g|_{1,T_0'}$. In particular, C does not depend on T' for T' in the range $[0,T_0']$.

But, v is defined from u via the transformation (3.3). Hence the desired result follows, and the corollary is established. □

Proof of Proposition 3.5. First note that since $f\in C_b^3(\overline{\mathbb{R}}^+)\cap H^1(\mathbb{R}^+)$, $f(x)$, $f'(x)$, $f''(x)\to 0$ as $x\to+\infty$ (cf. [9]). Let v be defined from u as in (3.3). Then by Lemma 3.4, $\partial_x^i\partial_t^j v(x,t)\to 0$, as $x\to+\infty$, uniformly for $0\le t\le T$, for $0\le j\le 1$ and $0\le i\le 2+j$. Because u is recovered from v by (3.35), $\partial_x^\nu\partial_t^\mu u(x,t)\to 0$, as $x\to+\infty$, uniformly for $0\le t\le T$, for μ and ν with $\mu+\nu\le 2$. Thus $u, u_x, u_t, u_{xx}, u_{xt}$ and u_{tt} tend to zero at $+\infty$, uniformly for $0\le t\le T$.

Let $U(x,t)=g(t)e^{-x}$ and $w=u-U$. There is a constant c_* such that, for $0\le t\le T$,

$$\|U(\cdot,t)\|\le|g(t)|\le c_*|g|_{1,T}.$$

So to prove (3.33), it is enough to establish a similar estimate for w. Now w satisfies the initial- and boundary-value problem

$$w_t+w_x+ww_x+w_{xxx}-\varepsilon w_{xxt}=\varphi-(wU)_x \quad \text{in } \overline{\mathbb{R}}^+\times[0,T], \tag{3.37}$$

where $\varphi=-(U_t+U_x+UU_x+U_{xxx}-\varepsilon U_{xxt})$, and

$$\begin{aligned} w(x,0)&=f(x)-g(0)e^{-x} && \text{for } x\in\overline{\mathbb{R}}^+,\\ w(0,t)&=0 && \text{for } t\in[0,T]. \end{aligned} \tag{3.38}$$

Multiply (3.37) by $2w$ and integrate the resulting expression over $(0,M)\times(0,t)$. There appears, after integrations by parts, and using the auxiliary conditions (3.38),

$$\begin{aligned} &\int_0^M\left[w^2(x,t)+\varepsilon w_x^2(x,t)\right]dx+\int_0^t w_x^2(0,s)\,ds\\ &\quad=\int_0^M\left[w^2(x,0)+\varepsilon w_x^2(x,0)\right]dx\\ &\qquad+\int_0^t\Big[-w^2(M,s)-\frac{2}{3}w^3(M,s)-2w(M,s)w_{xx}(M,s)\\ &\qquad\qquad+w_x^2(M,s)+2\varepsilon w(M,s)w_{xt}(M,s)-w^2(M,s)U(M,s)\Big]ds\\ &\qquad+2\int_0^t\int_0^M\varphi(x,s)w(x,s)\,dx\,ds-\int_0^t\int_0^M U_x(x,s)w^2(x,s)\,dx\,ds. \end{aligned} \tag{3.39}$$

Because $U(x,t)=g(t)e^{-x}$, it follows that

$$\|U\|_{C_b(\overline{\mathbb{R}}^+\times[0,T])},\ \|U_x\|_{C_b(\overline{\mathbb{R}}^+\times[0,T])}\le\|g\|_{C(0,T)}\le c_*|g|_{1,T}.$$

Similarly, since $\varepsilon\le 1$,

$$\|\varphi(\cdot,t)\|\le 2|g'(t)|+2|g(t)|+g^2(t),$$

so that

$$\int_0^t\int_0^M\varphi^2(x,s)\,dx\,ds\le C_1(|g|_{1,T}),$$

for all $(M,t)\in\overline{\mathbb{R}}^+\times[0,T]$. If

$$W_M(t)=\int_0^M\left[w^2(x,t)+\varepsilon w_x^2(x,t)\right]dx,$$

and if h_M denotes the supremum, over $[0,T]$, of the second integral on the right-hand side of (3.39), then the inequality

$$W_M(t)\le W_M(0)+h_M+C_1(|g|_{1,T})+c_*|g|_{1,T}\int_0^t W_M(s)\,ds$$

emerges. Gronwall's lemma implies

$$W_M(t)\le\big[W_M(0)+h_M+C_1(|g|_{1,T})\big]e^{c_*t|g|_{1,T}},$$

for $0\le t\le T$. Reference to (3.38) will convince the reader that $w(\cdot,0)\in H^1(\mathbb{R}^+)$. So $W_M(0)$ is bounded, as $M\to+\infty$. In fact,

$$W_M(0)\to\int_0^\infty\big[w^2(x,0)+\varepsilon w_x^2(x,0)\big]\,dx=W(0),$$

as $M\to+\infty$. Since u and u_x tend to zero as $x\to+\infty$, uniformly for $0\le t\le T$, so also do w and w_x. It follows that $h_M\to 0$ as $M\to+\infty$. Hence,

$$\overline{\lim_{M\to\infty}}\ W_M(t)\le\big[W(0)+C_1(|g|_{1,T})\big]e^{c_*T|g|_{1,T}},$$

for all $t\in[0,T]$. Thus for each $t\in[0,T]$, $w(\cdot,t)\in H^1(\mathbb{R}^+)$, and

$$\|w\|\le C_2\big(\|f\|+\varepsilon^{1/2}\|f_x\|,|g|_{1,T}\big),$$

for any ε in $(0,1]$. This is the desired bound on the $L^2(\mathbb{R}^+)$-norm of w, and so (3.33) is shown to be valid.

Now multiply the regularized equation (3.1a) by the combination $2\varepsilon u_{xt}-2u_{xx}-u^2$ and integrate the resulting relation over $\mathbb{R}^+\times(0,t)$. After integrations by parts, in which the fact that u and various of its derivatives vanish at $+\infty$ is used repeatedly, it is verified that

$$\begin{aligned}(3.40)\qquad &(1+\varepsilon)\int_0^\infty u_x^2(x,t)\,dx+\int_0^t\big[u_x^2(0,s)+H^2(s)\big]\,ds\\ &=(1+\varepsilon)\int_0^\infty f_x^2(x)\,dx-\frac{1}{3}\int_0^\infty f^3(x)\,dx\\ &\quad+\frac{1}{3}\int_0^\infty u^3(x,t)\,dx+\int_0^t\Big[\frac{1}{3}g^3(s)+\varepsilon g_t^2(s)\Big]\,ds-2\int_0^t g_t(s)u_x(0,s)\,ds,\end{aligned}$$

where

$$H(s)=u_{xx}(0,s)-\varepsilon u_{xt}(0,s)+\frac{1}{2}g^2(s).$$

Elementary inequalities, including (2.5), show that

$$\begin{aligned}\int_0^\infty u^3(x,t)\,dx&\le\|u(\cdot,t)\|^2\|u(\cdot,t)\|_{C_b(\overline{\mathbb{R}}^+)}\le\sqrt{2}\,\|u(\cdot,t)\|^{5/2}\|u_x(\cdot,t)\|^{1/2}\\ &\le\frac{1}{2}\|u_x(\cdot,t)\|^2+\|u(\cdot,t)\|^{10/3}.\end{aligned}$$

Putting together (3.40), the last observation, and the already established (3.33) yields,

$$\begin{aligned}&\|u_x(\cdot,t)\|^2+\int_0^t\big[u_x^2(0,s)+H^2(s)\big]\,ds\\ &\qquad\le 2a_0^{10/3}+2(1+\varepsilon)\|f_x\|^2-\frac{2}{3}\int_0^\infty f^3(x)\,dx+2\int_0^t\Big[-\frac{1}{3}g^3(s)+(1+\varepsilon)g_t^2(s)\Big]\,ds,\end{aligned}$$

where a_0 is the constant on the right of (3.33). Inequality (3.34) now follows, and the proposition is proved. □

A theorem of global existence of solutions of (3.1) and (3.4) is now in view. Its statement is postponed until after examination of one other aspect, of importance in the analysis in §§4 and 5. This aspect is embodied in the next proposition.

PROPOSITION 3.7. *Let* $F \in C_b^k(\overline{\mathbb{R}}^+) \cap H^k(\mathbb{R}^+)$ *and* $G \in C^m(0,T)$, *with* $F(0)=G(0)$, $k \geq 3$, $m \geq 1$ *and* $k \geq m$. *Let* v *be the solution of* (3.4) *defined in* $\mathcal{C}_T$. *Then there exists a constant* C *such that, for each* $t \in [0,T]$,

$$(*) \qquad \left\| \partial_x^i \partial_t^j v(\cdot,t) \right\|_{L^2((t,\infty))} \leq C,$$

provided that $0 \leq j \leq m$ *and* $0 \leq i \leq k+j$.

Proof. Throughout the demonstration, C will denote various constants which are independent of t in $[0,T]$. It will be convenient to introduce another condition, denoted $(*)_1$, which, for a function w defined on Ω_T, amounts to the requirement that $w(\cdot,t) \in H^1((t,\infty))$ for $t \in [0,T]$, and that

$$(*)_1 \qquad \|w(\cdot,t)\|_{H^1((t,\infty))} \leq C,$$

independently of t in $[0,T]$.

According to (3.33) and (3.34) in Lemma 3.5, $(*)_1$ holds for v itself. Thus v and v_x satisfy $(*)$. For one-dimensional domains, H^1 is an algebra, so that products of H^1 functions are again in H^1. Thus $(1+\varepsilon)v+\frac{1}{2}v^2$ satisfies $(*)_1$. Hence if, as before,

$$J(x,t)=\int_0^t \left[(1+\varepsilon)v(x,s)+\frac{1}{2}v^2(x,s) \right] ds,$$

then J satisfies $(*)_1$. So J and J_t satisfy $(*)_1$. It then follows from formula (3.21) that v_t satisfies $(*)_1$ as well. This observation may be used inductively to show that $\partial_t^i v$ satisfies $(*)_1$, for $0 \leq i \leq m$. Turning now to spatial derivatives, since $k>1$ formula (3.24) shows that v_x satisfies $(*)_1$. This means in particular that J_x satisfies $(*)_1$. Since $k>2$, then $F_{xx} \in H^1(\mathbb{R}^+)$, so, by reference to (3.26), one sees that v_{xx} satisfies $(*)_1$. Proceeding inductively, and using (3.27), it follows that $\partial_x^j v$ satisfies $(*)_1$ if $j \leq k-1$, and so $\partial_x^k v$ satisfies $(*)$.

From (3.28), v_{xt} is observed to satisfy $(*)_1$. The differential equation (3.4a) shows that v_{xxt} satisfies $(*)_1$. Using the differential equation, the results already in hand, and induction, mixed partial derivatives of the form $\partial_x^i \partial_t^j v$, where $j \geq 1$ and $i \geq 2$, are seen to satisfy $(*)_1$ when $j \leq m$ and $i \leq k+j-1$. Hence $\partial_x^i \partial_t^j v$ satisfies $(*)$ provided that $0 \leq j \leq m$ and $0 \leq i \leq k+j$. The desired results are now all established. □

It is worth summarizing the accomplishments of the present section. As the transformed problem (3.4) is only of transient interest, the theory is recapitulated in terms of the regularized problem (3.1). Thus the results stated now are consequences of the established propositions and the transformation (3.35) taking (3.4) to (3.1).

THEOREM 3.8. *Let* $\varepsilon>0$ *and* $T>0$ *be given. Suppose* $f \in C_b^k(\overline{\mathbb{R}}^+)$ *and* $g \in C^m(0,T)$ *with* $f(0)=g(0)$, $k \geq 3$, $m \geq 1$, *and* $k \geq m$. *Then there exists* $T_0>0$ *and a unique function* u *in* $C_b(\overline{\mathbb{R}}^+ \times [0,T_0])$ *which is a classical solution of the initial- and boundary-value problem* (3.1) *corresponding to the given* f *and* g. *Additionally,*

$$(3.41) \qquad \partial_x^i \partial_t^j u \in C_b(\overline{\mathbb{R}}^+ \times [0,T]),$$

for i *and* j *such that* $0 \leq j \leq m$, $0 \leq i \leq k$, *and* $i+j \leq k$. *Moreover, if* $f \in H^r(\mathbb{R}^+)$, *where* $r \geq 1$, *then* u *may be extended to a solution of* (3.1) *on* $\overline{\mathbb{R}}^+ \times [0,T]$. *In that case, there is a constant* C *such that, for* $0 \leq t \leq T$,

$$\left\| \partial_x^i \partial_x^j u(\cdot,t) \right\| \leq C,$$

for i *and* j *such that* $0 \leq j \leq \min\{r,m\}$, $0 \leq i \leq r$, *and* $i+j \leq r$.

As a corollary to this theorem, the following result emerges. It is this corollary which will find explicit use in the upcoming sections.

COROLLARY 3.9. *Let $\varepsilon>0$ be given. Let $f\in H^\infty(\mathbb{R}^+)$ and $g\in C^\infty(\mathbb{R}^+)$, with $f(0)=g(0)$. Then there exists a unique solution u of* (3.1) *defined on the quarter-plane $\overline{\mathbb{R}}^+\times\overline{\mathbb{R}}^+$ which is bounded on finite time intervals and which corresponds to the data f and g. Moreover, $u\in C^\infty(\overline{\mathbb{R}}^+\times\overline{\mathbb{R}}^+)$ and, for each $k\geq 0$,*

$$\partial_x^i\partial_t^j u\in C\big(\overline{\mathbb{R}}^+;H^k(\mathbb{R}^+)\big), \tag{3.42}$$

for all $i,j\geq 0$.

Proof. The existence of global solutions follows immediately from the theorem and the uniqueness result. Also, for any $i,j\geq 0$, $k>0$, and $T>0$, $w=\partial_x^i\partial_t^j u$ is uniformly bounded in $H^k(\mathbb{R}^+)$, for $0\leq t\leq T$.

It remains only to check that the mapping $t\to w(\cdot,t)$ is continuous, from $[0,T]$ to $H^k(\mathbb{R}^+)$. But, in fact, $u\in L^\infty(0,T;H^k(\mathbb{R}^+))$ and $u_t\in L^\infty(0,T;H^k(\mathbb{R}^+))$. It follows immediately (cf. [19]) that $u\in C(0,T;H^k(\mathbb{R}^+))$. The corollary is now verified. □

4. Estimates in $H^3(\mathbb{R}^+)$ for the regularized problem. The purpose of this and the next section is to derive a priori bounds, which do not depend on ε, for solutions of the regularized initial- and boundary-value problem,

$$u_t+u_x+uu_x+u_{xxx}-\varepsilon u_{xxt}=0 \quad \text{in } \overline{\mathbb{R}}^+\times[0,T], \tag{4.1a}$$

and

$$\begin{aligned} u(x,0)&=f(x) \quad \text{for } x\in\overline{\mathbb{R}}^+,\\ u(0,t)&=g(t) \quad \text{for } t\in[0,T]. \end{aligned} \tag{4.1b}$$

Here T is a fixed positive real number, and the aspired-for bounds will hold independently of t in $[0,T]$.

Throughout this section it will be assumed that $f\in H^\infty(\mathbb{R}^+)$, $g\in C^\infty(0,T)$, and $f(0)=g(0)$. In consequence of Corollary 3.9, for any ε in $(0,1]$, there is a classical solution $u=u_\varepsilon$ of (4.1) which is such that

$$u\in C^\infty\big(\overline{\mathbb{R}}^+\times[0,T]\big),$$

and, for integers $j,k\geq 0$,

$$\partial_t^j u\in C\big(0,T;H^k(\mathbb{R}^+)\big).$$

Some preliminary relations, established via energy arguments, will be derived in a sequence of technical lemmas. These prefatory results will be combined to obtain ε-independent bounds for u within the function class $C(0,T;H^3(\mathbb{R}^+))$ and for u_t within the function class $C(0,T;H^1(\mathbb{R}^+))$.

As a start on this program, recall that from Proposition 3.5, there is a constant a_1, depending only on $\|f\|_1$ and $|g|_{1,T}$, such that, independently of ε in $(0,1]$,

$$\|u(\cdot,t)\|_1^2+\int_0^t\big[u_x^2(0,s)+\big(u_{xx}(0,s)-\varepsilon u_{xt}(0,s)\big)^2\big]\,ds\leq a_1, \tag{4.2}$$

for all t in $[0,T]$. So, from (2.5) it follows that

$$\|u\|_{C_b(\overline{\mathbb{R}}^+\times[0,T])}^2\leq 2\sup_{0\leq t\leq T}\big\{\|u_x(\cdot,t)\|\,\|u(\cdot,t)\|\big\}\leq a_1, \tag{4.3}$$

and, because of the differential equation (4.1a),

$$\begin{aligned} \int_0^t\big(u_{xxx}(0,s)-\varepsilon u_{xxt}(0,s)\big)^2ds&=\int_0^t\big(g_t(s)+u_x(0,s)+g(s)u_x(0,s)\big)^2ds\\ &\leq c=c\big(\|f\|_1,|g|_{1,T}\big), \end{aligned} \tag{4.4}$$

for all t in $[0,T]$.

If u is the solution of (4.1) and $t\in[0,T]$, define

$$A^2(t)=\sup_{0\le s\le t}\left\{\|u(\cdot,s)\|_3^2+\varepsilon\|u_{xxxx}(\cdot,s)\|^2\right\}$$
$$+\int_0^t\left[u_{xxxx}^2(0,s)+u_{xxx}^2(0,s)+u_{xx}^2(0,s)+\varepsilon^2u_{xt}^2(0,s)+\varepsilon u_{xxt}^2(0,s)\right]ds,$$

and

$$B^2(t)=\sup_{0\le s\le t}\|u_t(\cdot,s)\|_1^2+\int_0^t u_{xt}^2(0,s)\,ds.$$

It will be shown that $A(t)$ and $B(t)$ are bounded on $[0,T]$, independently of ε small enough. The first step in obtaining this result is the following $H^2(\mathbb{R}^+)$-estimate.

LEMMA 4.1. *Let $T>0$, $f\in H^\infty(\mathbb{R}^+)$, $g\in H^\infty(0,T)$, with $f(0)=g(0)$. There exist positive constants ε_1, a_2 and c_1, where*

$$\varepsilon_1=\varepsilon_1\left(\|f\|_1,|g|_{1,T}\right),\qquad a_2=a_2\left(\|f\|_2+\varepsilon_1^{1/2}\|f_{xxx}\|,|g|_{1,T}\right),$$
$$c_1=c_1\left(\|f\|_1,|g|_{1,T}\right),$$

such that the solution u of (4.1) corresponding to the data f and g satisfies

$$\|u(\cdot,t)\|_2^2+\int_0^t\left[u_{xxx}^2(0,s)+u_{xx}^2(0,s)+\varepsilon^2u_{xt}^2(0,s)\right]ds$$
$$\le a_2+c_1\varepsilon\int_0^t A^2(s)B(s)\,ds-\frac{18}{5}\int_0^t u_{xx}(0,s)u_{xt}(0,s)\,ds,$$

provided that $t\in[0,T]$ and $\varepsilon\in(0,\varepsilon_1]$.

Remark. The presence of the last term on the right-hand side of the above inequality means that this estimate is not directly effective in bounding $\|u(\cdot,t)\|_2$, independently of ε.

Proof. For each t in $[0,T]$, define $V(t)$ as

$$V(t)=\int_0^\infty\left[\left(\frac{9}{5}-3\varepsilon u\right)u_{xx}^2-3uu_x^2+\frac{1}{4}u^4+\frac{9}{5}\varepsilon u_{xxx}^2\right]dx.$$

Multiply (4.1a) by $u^3-3u_x^2$, differentiate (4.1a) once with respect to x and multiply the result by $-6uu_x-\frac{18}{5}u_{xxx}$, add the two equations thus obtained, and integrate their sum over $\mathbb{R}^+\times(0,t)$. After several integrations by parts, there appears,

(4.5)

$$V(t)-V(0)+\frac{9}{5}\int_0^t\left[u_{xxx}^2(0,s)+u_{xx}^2(0,s)\right]ds$$
$$=\int_0^t\Big[\frac{1}{4}g^4(s)+\frac{1}{5}g^5(s)-3g(s)u_x^2(0,s)+g^3(s)u_{xx}(0,s)-\frac{9}{2}g^2(s)u_x^2(0,s)$$
$$-6g(s)u_x(0,s)u_{xxx}(0,s)+\frac{6}{5}g(s)u_{xx}^2(0,s)$$
$$-\frac{3}{5}u_x^2(0,s)u_{xx}(0,s)-\frac{18}{5}u_{xx}(0,s)u_{xt}(0,s)\Big]ds$$
$$+\varepsilon\int_0^t\left[6g_t(s)u_x(0,s)u_{xx}(0,s)+6g(s)u_x(0,s)u_{xxt}(0,s)\right.$$
$$\left.-3u_x^2(0,s)u_{xt}(0,s)-g^3(s)u_{xt}(0,s)\right]ds$$
$$+\varepsilon\int_0^t\int_0^\infty\left[3u_{xx}^2u_t+6u_tu_xu_{xxx}-3u^2u_xu_{xt}\right]dx\,ds.$$

Because of the relation (4.2), the first seven boundary terms on the right-hand side of (4.5) can be bounded in terms of the data f and g and a suitable small multiple of the two boundary integrals on the left-hand side of (4.5). Using (2.5) and (4.2), it follows that for any $\delta>0$,

$$
\begin{aligned}
(4.6)\quad & \int_0^t u_x^2(0,s)u_{xx}(0,s)\,ds \\
& \le \|u_x\|_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\left(\int_0^t u_x^2(0,s)\,ds\int_0^t u_{xx}^2(0,s)\,ds\right)^{1/2} \\
& \le \sqrt{2}\left\{\sup_{0\le s\le t}\left(\|u_x(\cdot,s)\|^{1/2}\|u_{xx}(\cdot,s)\|^{1/2}\right)\right. \\
& \qquad \left.\cdot\left(\int_0^t u_x^2(0,s)\,ds\int_0^t u_{xx}^2(0,s)\,ds\right)^{1/2}\right\} \\
& \le a_1^3\delta^{-3}+\delta\left\{\sup_{0\le s\le t}\|u_{xx}(\cdot,s)\|^2+\int_0^t u_{xx}^2(0,s)\,ds\right\}.
\end{aligned}
$$

Since

$$
\int_0^t g_t(s)u_x(0,s)u_{xx}(0,s)\,ds\le\|u_x\|_{C_b(\bar{\mathbb{R}}^+\times[0,t])}|g_t|_T\left(\int_0^t u_{xx}^2(0,s)\,ds\right)^{1/2},
$$

a similar bound holds for the term

$$
\varepsilon\int_0^t g_t(s)u_x(0,s)u_{xx}(0,s)\,ds.
$$

The estimate (4.2) also implies that

$$
\begin{aligned}
\varepsilon^2\int_0^t u_{xt}^2(0,s)\,ds &\le 2\left\{\int_0^t(u_{xx}(0,s)-\varepsilon u_{xt}(0,s))^2ds+\int_0^t u_{xx}^2(0,s)\,ds\right\} \\
&\le 2a_1+2\int_0^t u_{xx}^2(0,s)\,ds.
\end{aligned}
$$

As a consequence, bounds similar to that in (4.6) obtain for the terms

$$
\varepsilon\int_0^t u_x^2(0,s)u_{xt}(0,s)\,ds \quad\text{and}\quad \varepsilon\int_0^t g^3(s)u_{xt}(0,s)\,ds.
$$

Making use of (4.4), the term,

$$
\varepsilon\int_0^t g(s)u_x(0,s)u_{xxt}(0,s)\,ds,
$$

may be bounded in the same way.

Still relying on (4.2) and (4.3), the term

$$
3\int_0^\infty uu_x^2\,dx\le 3\|u\|_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\int_0^\infty u_x^2\,dx\le 3a_1^{3/2}.
$$

Hence,

$$
\int_0^\infty\left(\frac{9}{5}-3\varepsilon u\right)u_{xx}^2\,dx\le V(t)+3a_1^{3/2}.
$$

But, by (4.3), $|u|$ does not exceed the value $a_1^{1/2}$ on $\overline{\mathbb{R}}^+\times[0,T]$. Consequently, if $\varepsilon_1=(25a_1)^{-1/2}$, then for $0<\varepsilon\le\varepsilon_1$,

$$\frac{6}{5}\int_0^\infty u_{xx}^2\,dx\le V(t)+3a_1^{3/2},$$

for all t in $[0,T]$.

Therefore, if (4.5) and a suitable multiple of (4.2) are summed, and use is made of the above estimates, then for t in $[0,T]$ and ε in $(0,\varepsilon_1]$,

$$\|u(\cdot,t)\|_2^2+\int_0^t\left[u_{xxx}^2(0,s)+u_{xx}^2(0,s)+\varepsilon^2u_{xt}^2(0,s)\right]ds$$

$$\le a_2-\frac{18}{5}\int_0^t u_{xx}(0,s)u_{xt}(0,s)\,ds$$

$$+\varepsilon\int_0^t\int_0^\infty\left[3u_{xx}^2u_t+6u_xu_{xxx}u_t-3u^2u_xu_{xt}\right]dx\,ds.$$

Here, the constant a_2 stems from $V(0)$ and from the various combinations of a_1 that appear in the foregoing estimates. The desired result now follows from the last relation, (4.2), and the definitions of $A(t)$ and $B(t)$. □

The estimate of the $H^2(\mathbb{R}^+)$-norm of the solution u of (4.1) given in Lemma 4.1 will be used in determining the following bound for $A(t)$.

LEMMA 4.2. *Let $T>0$, $f\in H^\infty(\mathbb{R}^+)$, $g\in C^\infty(0,T)$, with $f(0)=g(0)$. There exist positive constants a_3 and c_2, where*

$$a_3=a_3\left(\|f\|_3+\varepsilon_1^{1/2}\|f_{xxxx}\|,|g|_{2,T}\right)\quad\text{and}\quad c_2=c_2\left(\|f\|_1,|g|_{1,T}\right),$$

such that the solution of (4.1) *corresponding to f and g satisfies*

$$A^2(t)-\varepsilon c_2\left[A^3(t)+\varepsilon\left(1+A^2(t)\right)B^2(t)\right]\le a_3+\varepsilon^{1/2}c_2\int_0^t A^2(s)B(s)\,ds,$$

for all t in $[0,T]$ and ε in $(0,\varepsilon_1]$.

Remark. The ε_1 appearing in the above statement is that derived already in Lemma 4.1.

Proof. As in the proof of the last lemma, the desired result will be obtained from a technical "energy" argument. In the proof, various constants dependent on aspects of the data f and g will appear. These will generally be denoted simply by c, and this symbol's occurrence in different formulae is not taken to connote the same constant. Define, for each t in $[0,T]$,

$$W(t)=\int_0^\infty\left[\frac{108}{35}\left(\varepsilon u_{xxxx}^2+u_{xxx}^2\right)-\frac{36}{5}\left(u-\varepsilon u_{xx}\right)u_{xx}^2\right.$$

$$\left.+6\left(u_x^2+\varepsilon u_{xx}^2\right)-\frac{1}{5}u^5-3\varepsilon u_x^4-\frac{36}{5}\varepsilon uu_{xxx}^2\right]dx.$$

Multiply (4.1a) by $12uu_x^2-\frac{36}{5}u_{xx}^2-u^4$, differentiate (4.1a) once with respect to x and multiply this by $12u^2u_x$, differentiate (4.1a) twice with respect to x and multiply this by $-\frac{216}{35}u_{xxxx}-\frac{72}{5}uu_{xx}$, add the three resulting equations and integrate their sum over

$\mathbb{R}^+\times(0,t)$. After many integrations by parts with respect to the spatial variable x, there appears,

(4.7)

$$
\begin{aligned}
W(t)-W(0)+\frac{108}{35}\int_0^t\left[u_{xxxx}^2(0,s)+u_{xxx}^2(0,s)\right]ds
&=\int_0^t\Big[-\frac{36}{5}g(s)u_{xx}^2(0,s)+6g^2(s)u_x^2(0,s)-\frac{1}{5}g^5(s)-\frac{1}{6}g^6(s)\\
&\quad-g^4(s)(u_{xx}(0,s)-\varepsilon u_{xt}(0,s))+8g^3(s)u_x^2(0,s)\\
&\quad+12g^2(s)u_x(0,s)(u_{xxx}(0,s)-\varepsilon u_{xxt}(0,s))\\
&\quad-12g(s)u_x^2(0,s)(u_{xx}(0,s)-\varepsilon u_{xt}(0,s))+3u_x^4(0,s)-\frac{66}{5}g^2(s)u_{xx}^2(0,s)\\
&\quad+\frac{72}{5}u_x(0,s)u_{xx}(0,s)(u_{xxx}(0,s)-\varepsilon u_{xxt}(0,s))\\
&\quad-\frac{72}{5}u_x(0,s)u_{xxx}(0,s)(u_{xx}(0,s)-\varepsilon u_{xt}(0,s))\\
&\quad-\frac{144}{35}u_x(0,s)u_{xx}(0,s)u_{xxx}(0,s)\\
&\quad+\frac{144}{35}g(s)u_{xxx}^2(0,s)-\frac{72}{5}g(s)u_{xx}(0,s)u_{xxxx}(0,s)-\frac{36}{35}u_{xx}^3(0,s)\\
&\quad-\frac{216}{35}u_{xxx}(0,s)u_{xxt}(0,s)+\frac{72}{5}\varepsilon g(s)u_{xx}(0,s)u_{xxxt}(0,s)\Big]ds\\
&\quad+\varepsilon\int_0^t\int_0^\infty\Big[4u^3u_xu_{xt}+24uu_xu_{xx}u_{xt}+\frac{72}{5}u_{xx}u_{xxx}u_{xt}\\
&\quad+\frac{72}{5}u_xu_{xxxx}u_{xt}+12uu_tu_{xx}^2-\frac{36}{5}u_tu_{xxx}^2\Big]dx\,ds.
\end{aligned}
$$

First note that, because of (4.2), there is a positive constant c, depending on $\|f\|_1$ and $|g|_{1,T}$, so that

$$
\frac{108}{35}\left(\|u_{xxx}(\cdot,t)\|^2+\varepsilon\|u_{xxxx}(\cdot,t)\|^2\right)-\frac{72}{5}\varepsilon A^3(t)\le W(t)+c,
$$

for all t in $[0,T]$. Also, in consequence of (2.1) and (4.2), there is another constant c, depending again on $\|f\|_1$ and $|g|_{1,T}$, such that, for any $\delta>0$,

(4.8)
$$
\begin{aligned}
\|u_x\|^2_{C_b(\overline{\mathbb{R}}^+\times[0,t])}&\le 2\left\{\sup_{0\le s\le t}\left(\|u_x(\cdot,s)\|\|u_{xx}(\cdot,s)\|\right)\right\}\\
&\le c\delta^{-1}+\delta\left\{\sup_{0\le s\le t}\|u_{xx}(\cdot,s)\|^2\right\}.
\end{aligned}
$$

By an analogous argument,

(4.9)
$$
\|u_{xx}\|^2_{C_b(\overline{\mathbb{R}}^+\times[0,t])}\le c\delta^{-3}+\delta\left\{\sup_{0\le s\le t}\|u_{xxx}(\cdot,s)\|^2\right\}.
$$

Taken together with (4.2), these estimates imply that there is a constant c, depending on $\|f\|_1$ and $|g|_{1,T}$, such that for all $\delta>0$,

$$
\begin{aligned}
(4.10)\qquad & \int_0^t u_x(0,s)u_{xx}(0,s)u_{xxx}(0,s)\,ds \\
&\le \|u_{xx}\|_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\left[\int_0^t u_x^2(0,s)\,ds\int_0^t u_{xxx}^2(0,s)\,ds\right]^{1/2} \\
&\le c\left[\delta^{-3}+\int_0^t u_{xxx}^2(0,s)\,ds\right]+\delta\left\{\sup_{0\le s\le t}\|u_{xxx}(\cdot,s)\|^2\right\}.
\end{aligned}
$$

By adding (4.7) and a suitable positive multiple $\alpha=\alpha(\sup_{0\le s\le t}\|u(\cdot,s)\|_1)$ of the inequality stated in Lemma 4.1, and using (4.2) and (4.3), bounds similar to those exhibited in (4.10) may be shown to hold for all the boundary terms on the right-hand side of (4.7) except for the last three. Choosing δ appropriately, it may thus be inferred that, for all ε in $(0,\varepsilon_1]$, and for all t in $[0,T]$,

$$
\begin{aligned}
(4.11)\qquad & 3\left[A^2(t)-\varepsilon\int_0^t u_{xxt}^2(0,s)\,ds\right]-\frac{72}{5}\varepsilon A^3(t) \\
&\le \tilde{a}_3+\tilde{c}_2\varepsilon^{1/2}\int_0^t A^2(s)B(s)\,ds \\
&\quad -\int_0^t\left[\frac{36}{35}u_{xx}^3(0,s)+\frac{216}{35}u_{xxx}(0,s)u_{xxt}(0,s)\right. \\
&\qquad \left.-\frac{72}{5}\varepsilon g(s)u_{xx}(0,s)u_{xxxt}(0,s)+\frac{18}{5}\alpha u_{xx}(0,s)u_{xt}(0,s)\right]ds.
\end{aligned}
$$

Here $\tilde{a}_3=\tilde{a}_3(\|f\|_3+\varepsilon_1^{1/2}\|f_{xxxx}\|,|g|_{2,T})$ and $\tilde{c}_2=\tilde{c}_2(\|f\|_1,|g|_{1,T})$.

To complete the proof of the lemma, it suffices to control suitably the boundary terms appearing on the right side of inequality (4.11). To this end, observe first that (4.2) and (4.9) imply

$$
\begin{aligned}
(4.12)\qquad \int_0^t u_{xx}^3(0,s)\,ds &\le \|u_{xx}\|^2_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\int_0^t|u_{xx}(0,s)|\,ds \\
&\le \|u_{xx}\|^2_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\int_0^t\big(|u_{xx}(0,s)-\varepsilon u_{xt}(0,s)|+\varepsilon|u_{xt}(0,s)|\big)\,ds \\
&\le c\big(\delta^{-3}+\delta A^2(t)\big)\big(1+\varepsilon^2B^2(t)\big),
\end{aligned}
$$

for any $\delta>0$, where the constant c depends on $\|f\|_1$, $|g|_{1,T}$ and T. Next note that equation (4.1a) implies

$$
\begin{aligned}
&-\int_0^t u_{xxx}(0,s)u_{xxt}(0,s)\,ds \\
&\qquad =\int_0^t\left[g_t(s)+u_x(0,s)+g(s)u_x(0,s)-\varepsilon u_{xxt}(0,s)\right]u_{xxt}(0,s)\,ds.
\end{aligned}
$$

Integration by parts in the temporal variable yields

$$
\begin{aligned}
&\int_0^t\left[g_t(s)+u_x(0,s)+g(s)u_x(0,s)\right]u_{xxt}(0,s)\,ds \\
&\qquad =\left[g_t(s)+u_x(0,s)+g(s)u_x(0,s)\right]u_{xx}(0,s)\Big|_{s=0}^{s=t} \\
&\qquad\quad -\int_0^t\left[g_{tt}(s)+u_{xt}(0,s)+g_t(s)u_x(0,s)+g(s)u_{xt}(0,s)\right]u_{xx}(0,s)\,ds.
\end{aligned}
$$

From (4.1a) it also follows that

$$u_{xt}(0,s)=\varepsilon u_{xxxt}(0,s)-\left[u_{xx}(0,s)+u_x^2(0,s)+g(s)u_{xx}(0,s)+u_{xxxx}(0,s)\right]. \tag{4.13}$$

Hence, due to (4.8) and (4.9), for any $\delta>0$ there is a constant c_δ, depending on δ, $\|f\|_1$ and $|g|_{2,T}$, such that

$$\begin{aligned}-\int_0^t u_{xxx}(0,s)u_{xxt}(0,s)\,ds &\\ \le c_\delta-\varepsilon\int_0^t u_{xxt}^2(0,s)\,ds &\\ +\delta\left[A^2(t)+\int_0^t u_{xxxx}^2(0,s)\,ds\right]-\varepsilon\int_0^t(1+g(s))u_{xx}(0,s)u_{xxxt}(0,s)\,ds.&\end{aligned} \tag{4.14}$$

Similarly, it follows from (4.8), (4.9) and (4.13) that, for any $\delta>0$,

$$\begin{aligned}-\int_0^t u_{xx}(0,s)u_{xt}(0,s)\,ds &\\ \le c_\delta+\delta\left[A^2(t)+\int_0^t u_{xxxx}^2(0,s)\,ds\right]-\varepsilon\int_0^t u_{xx}(0,s)u_{xxxt}(0,s)\,ds,&\end{aligned} \tag{4.15}$$

where the constant c_δ depends on δ, $\|f\|_1$ and $|g|_{1,T}$. Combining (4.15) with (4.11), (4.12) and (4.14), and choosing δ in a perspicuous way, there appears,

$$\begin{aligned}2A^2(t)-\varepsilon\hat{c}_2\left[A^3(t)+\varepsilon\left(1+A^2(t)\right)B^2(t)\right] &\\ \le\hat{a}_3+\varepsilon^{1/2}\hat{c}_2\int_0^t A^2(s)B(s)\,ds &\\ -\varepsilon\int_0^t\left[\left(\frac{18}{5}\alpha+\frac{216}{35}\right)-\frac{288}{35}g(s)\right]u_{xx}(0,s)u_{xxt}(0,s)\,ds,&\end{aligned} \tag{4.16}$$

holding for all ε in $(0,\varepsilon_1]$ and t in $[0,T]$. Here,

$$\hat{a}_3=\hat{a}_3\left(\|f\|_3+\varepsilon_1^{1/2}\|f_{xxx}\|,|g|_{2,T}\right)\quad\text{and}\quad\hat{c}_2=\hat{c}_2\left(\|f\|_1,|g|_{1,T}\right).$$

To estimate the boundary terms on the right-hand side of (4.16), use (4.9) again to deduce that, corresponding to any $\delta>0$ there is another constant c_δ, dependent on δ, $\|f\|_1$ and $|g|_{1,T}$, such that

$$\begin{aligned}-\varepsilon\int_0^t\left[\left(\frac{18}{5}\alpha+\frac{216}{35}\right)-\frac{288}{35}g(s)\right]u_{xx}(0,s)u_{xxxt}(0,s)\,ds &\\ \le\delta^{-1}\left[\left(\frac{18}{5}\alpha+\frac{216}{35}\right)+\frac{288}{35}\|g\|_{C(0,T)}\right]^2\int_0^t u_{xx}^2(0,s)\,ds+\delta\varepsilon^2\int_0^t u_{xxxt}^2(0,s)\,ds &\\ \le c_\delta+\delta A^2(t)+\delta\varepsilon^2\int_0^t u_{xxxt}^2(0,s)\,ds.&\end{aligned} \tag{4.17}$$

So, the only term still presenting difficulty is the final one in (4.17). To estimate this quantity, differentiate the regularized equation (4.1a) twice with respect to x, multiply the result by $2\varepsilon u_{xxxt}$ and integrate over $\mathbb{R}^+\times(0,t)$. The effect of these operations is to produce the relation

$$\begin{aligned}\varepsilon\left(\|u_{xxx}(\cdot,t)\|^2-\|u_{xxxx}(\cdot,t)\|^2\right)+\varepsilon^2\int_0^t u_{xxxt}^2(0,s)\,ds &\\ =\varepsilon\left(\|f_{xxx}\|^2-\|f_{xxxx}\|^2\right)+\varepsilon\int_0^t u_{xxt}^2(0,s)\,ds &\\ +2\varepsilon\int_0^t u_{xxxx}(0,s)u_{xxxt}(0,s)\,ds-2\varepsilon\int_0^t\int_0^\infty(uu_x)_{xx}u_{xxxt}\,dx\,ds.&\end{aligned} \tag{4.18}$$

The last integral on the right-hand side of (4.18) seems somewhat awkward. However, after integration by parts,

$$\begin{aligned}\int_0^t\int_0^\infty (uu_x)_{xx}u_{xxxt}\,dx\,ds &= \int_0^t\int_0^\infty (3u_xu_{xx}+uu_{xxx})u_{xxxt}\,dx\,ds\\ &= \int_0^\infty\Big[\frac{1}{2}u(x,s)u_{xxx}^2(x,s)-u_{xx}^3(x,s)\Big]\,dx\Big|_{s=0}^{s=t}\\ &\quad +3\int_0^t u_x(0,s)\big[u_{xxx}(0,s)u_{xt}(0,s)-u_{xx}(0,s)u_{xxt}(0,s)\big]\,ds\\ &\quad +\int_0^t\int_0^\infty\Big(3u_{xx}u_{xxx}u_{xt}+3u_xu_{xxxx}u_{xt}-\frac{1}{2}u_{xxx}^2u_t\Big)\,dx\,ds.\end{aligned}$$

Also, by (4.8) there is a constant c, dependent on $\|f\|_1$ and $|g|_{1,T}$, such that

$$\begin{aligned}\varepsilon\int_0^t u_x(0,s)u_{xxx}(0,s)u_{xt}(0,s)\,ds &\le \varepsilon^2B^2(t)\|u_x\|^2_{C_b(\bar{\mathbb{R}}^+\times[0,t])}+\int_0^t u_{xxx}^2(0,s)\,ds\\ &\le c\varepsilon^2\big(1+A^2(t)\big)B^2(t)+A^2(t).\end{aligned}$$

And,

$$\begin{aligned}\varepsilon\int_0^t u_x(0,s)u_{xx}(0,s)u_{xxt}(0,s)\,ds &\le \varepsilon\|u_x\|^2_{C_b(\bar{\mathbb{R}}^+\times[0,t])}\int_0^t u_{xx}^2(0,s)\,ds+\varepsilon\int_0^t u_{xxt}^2(0,s)\,ds\\ &\le c\varepsilon A^3(t)+A^2(t).\end{aligned}$$

Referring to the definition of A and B below (4.4), and applying elementary estimates, it follows at once that

$$\begin{aligned}&2\varepsilon\int_0^t\int_0^\infty\Big(3u_{xx}u_{xxx}u_{xt}+3u_xu_{xxxx}u_{xt}-\frac{1}{2}u_{xxx}^2u_t\Big)\,dx\,ds\\ &\qquad\le\int_0^t\big[6\varepsilon A^2(s)B(s)+6\varepsilon^{1/2}A^2(s)B(s)+\varepsilon A^2(s)B(s)\big]\,ds\\ &\qquad\le 13\varepsilon^{1/2}\int_0^t A^2(s)B(s)\,ds.\end{aligned}$$

Here, and above, the restriction $\varepsilon\le 1$ is used. The last few relations combine with (4.18) to produce the inequality

$$\begin{aligned}(4.19)\qquad \varepsilon^2\int_0^t u_{xxxt}^2(0,s)\,ds &\le c+c'\varepsilon^{1/2}\int_0^t A^2(s)B(s)\,ds\\ &\quad +c''\big\{A^2(t)+\varepsilon\big[A^3(t)+\varepsilon\big(1+A^2(t)\big)B^2(t)\big]\big\}.\end{aligned}$$

If, in (4.17), δ is now chosen small enough, the desired inequality follows from (4.16), (4.17) and (4.19). This completes the proof of Lemma 4.2. □

To make effective use of Lemma 4.2, an estimate for $B(t)$ is needed. The following result will be sufficient.

LEMMA 4.3. *Let $T>0$, $f\in H^\infty(\mathbb{R}^+)$, $g\in H^\infty(0,T)$, with $f(0)=g(0)$. There are positive constants a_4 and c_3, with*

$$a_4=a_4\big(\|u_t(\cdot,0)\|_1,|g|_{2,T}\big)\quad\text{and}\quad c_3=c_3\big(\|f\|_3,|g|_{1,T}\big),$$

such that the solution of (4.1) *corresponding to the data f and g satisfies the inequality*

$$B^2(t)\le a_4+c_3\int_0^t\big[(1+A(s))B^2(s)+\varepsilon B^3(s)\big]\,ds,$$

for all t in $[0,T]$ and ε in $(0,1]$.

Proof. Let $v(x,t)=u_t(x,t)$. Then v satisfies the variable-coefficient partial differential equation

$$v_t+v_x+(uv)_x+v_{xxx}-\varepsilon v_{xxt}=0, \tag{4.20}$$

holding for (x,t) in $\overline{\mathbb{R}}^+\times[0,T]$. Multiply (4.20) by $2v$ and integrate over $\mathbb{R}^+\times(0,t)$, where $t\in[0,T]$. Then, it follows that

$$\begin{aligned}\|v(\cdot,t)\|^2+\varepsilon\|v_x(\cdot,t)\|^2+\int_0^t v_x^2(0,s)\,ds&\\ =\|v(\cdot,0)\|^2+\varepsilon\|v_x(\cdot,0)\|^2+\int_0^t(1+g(s))g_t^2(s)\,ds&\\ +2\int_0^t g_t(s)\big[v_{xx}(0,s)-\varepsilon v_{xt}(0,s)\big]\,ds-\int_0^t\int_0^\infty u_x v^2\,dx\,ds.&\end{aligned} \tag{4.21}$$

Next, multiply (4.20) by $2(\varepsilon v_{xt}-uv-v_{xx})$ and integrate again over $\mathbb{R}^+\times(0,t)$. This leads to

$$\begin{aligned}(1+\varepsilon)\|v_x(\cdot,t)\|^2-\int_0^\infty u(x,t)v^2(x,t)\,dx+\int_0^t\big\{v_x^2(0,s)+\big[v_{xx}(0,s)-\varepsilon v_{xt}(0,s)\big]^2\big\}\,ds&\\ =(1+\varepsilon)\|v_x(\cdot,0)\|^2-\int_0^\infty f(x)v^2(x,0)\,dx+\int_0^t g_{tt}^2(s)\big[\varepsilon-g^2(s)\big]\,ds&\\ -2\int_0^t g_{tt}(s)v_x(0,s)\,ds-2\int_0^t g(s)g_t(s)\big[v_{xx}(0,s)-\varepsilon v_{xt}(0,s)\big]\,ds&\\ +\int_0^t\int_0^\infty(2uvv_x-v^3)\,dx\,ds.&\end{aligned} \tag{4.22}$$

The underlying equation (4.1a) implies that

$$-\int_0^t\int_0^\infty v^3\,dx\,dx=\int_0^t\int_0^\infty v^2(u_x+uu_x+u_{xxx}-\varepsilon u_{xxt})\,dx\,ds.$$

The last term on the right side of this relation is potentially troublesome, but after integration by parts,

$$\varepsilon\int_0^t\int_0^\infty v^2u_{xxt}\,dx\,dx=-\varepsilon\int_0^t g_t^2(s)v_x(0,s)\,ds-2\varepsilon\int_0^t\int_0^\infty vv_x^2\,dx\,ds.$$

Also,

$$\int_0^\infty u(x,t)v^2(x,t)\,dx\le\|u\|_{C_b(\overline{\mathbb{R}}^+\times[0,t])}\|v(\cdot,t)\|^2\le c\|v(\cdot,t)\|^2,$$

where c depends on $\|f\|_1$ and $|g|_{1,T}$, as in (4.3). The desired result thus follows by adding an appropriate multiple of (4.21) to (4.22) and making the kind of estimates based on (4.2) that are, by now, familiar. □

Recapitulating the outcome of Lemmas 4.2 and 4.3, if u is the solution of (4.1) corresponding to initial data f and boundary data g, and A and B are the associated functionals defined below (4.4), then A and B are restricted by the system of inequalities

$$\begin{aligned}A^2(t)-\varepsilon c_2\big[A^3(t)+\varepsilon(1+A^2(t))B^2(t)\big]\le a_3+\varepsilon^{1/2}c_2\int_0^t A^2(s)B(s)\,ds,&\\ B^2(t)\le a_4+c_3\int_0^t\big[(1+A(s))B^2(s)+\varepsilon B^3(s)\big]\,ds,&\end{aligned} \tag{4.23}$$

holding for all t in $[0,T]$ and ε in $(0,\varepsilon_1]$. The constants ε_1, a_3, a_4, c_2 and c_3 have all been previously determined to depend simply on T, on various norms of f and g and on $\|u_t(\cdot,0)\|_1$. The system (4.23) will be exploited to obtain the following bound on u, which holds uniformly for ε sufficiently small.

LEMMA 4.4. *Let $T>0$, $f\in H^\infty(\mathbb{R}^+)$, $g\in H^\infty(0,T)$ be given with $f(0)=g(0)$. Let u be the solution of* (4.1) *corresponding to the data f and g. There are positive constants ε_2 and c_4, both depending on $\|f\|_4$, $|g|_{2,T}$ and $\|u_t(\cdot,0)\|_1$, such that for ε in $(0,\varepsilon_2]$ and t in $[0,T]$, both $A(t)$ and $B(t)$ are no larger than c_4.*

Proof. For each $M\in\mathbb{R}$ such that

$$M>\max(A(0),B(0)), \tag{4.24}$$

let

$$t_M=\inf\{t\in[0,T]: A(t)\geq M \text{ or } B(t)\geq M\},$$

with the understanding that if the set over which the infimum is taken is empty, then $t_M=T$. To establish the lemma, it suffices to show that $t_M=T$ for some M and all sufficiently small ε.

Observe that on the interval $[0,t_M)$, where M is supposed chosen as above, (4.23) implies that

$$\begin{aligned}&[1-\varepsilon c_2M(1+\varepsilon M)]A^2(t)\leq a_3+\varepsilon^{1/2}c_2\int_0^t A^2(s)B(s)\,ds+c_2(\varepsilon M)^2,\\ &B^2(t)\leq a_4+c_3\int_0^t(1+A(s))B^2(s)\,ds+\varepsilon c_3TM^3.\end{aligned} \tag{4.25}$$

For each M satisfying (4.24), choose $\varepsilon_2=\varepsilon_2(M)\in(0,\min(\frac{1}{2},\varepsilon_1))$ such that for all ε in $(0,\varepsilon_2)$,

$$1-c_2\varepsilon M(1+\varepsilon M)\geq\frac{1}{2},\quad c_2(\varepsilon M)^2\leq 1,\quad c_3\varepsilon TM^3\leq 1. \tag{4.26}$$

Further, let $A_1(t)=1+A(t)$. Then from (4.25), it follows that for all t in $[0,t_M)$ and for all ε in $(0,\varepsilon_2)$,

$$\begin{aligned}A_1^2(t)&\leq 6+4a_3+4c_2\varepsilon^{1/2}\int_0^t A_1^2(s)B(s)\,ds,\\ B^2(t)&\leq 1+a_4+c_3\int_0^t A_1(s)B^2(s)\,ds.\end{aligned}$$

Hence, in this range of t and ε, there are positive constants α,β and γ, independent of M, such that

$$\begin{aligned}A_1^2(t)&\leq\frac{\alpha}{1-\varepsilon^{1/2}}+2\frac{\gamma}{\beta}\varepsilon^{1/2}\int_0^t A_1^2(s)B(s)\,ds,\\ B^2(t)&\leq\frac{\beta}{1-\varepsilon^{1/2}}+2\frac{\gamma}{\alpha}\int_0^t A_1(s)B^2(s)\,ds.\end{aligned} \tag{4.27}$$

(First choose α and β large enough, and then choose γ large enough. Note then that α,β and γ only depend on the constants a_3, a_4, c_2 and c_3.) Define $\bar{A}_1$ and $\bar{B}$ to be the maximal solution of the system

$$\begin{aligned}\bar{A}_1^2(t)&=\frac{\alpha}{1-\varepsilon^{1/2}}+2\frac{\gamma}{\beta}\varepsilon^{1/2}\int_0^t\bar{A}_1^2(s)\bar{B}(s)\,ds,\\ \bar{B}^2(t)&=\frac{\beta}{1-\varepsilon^{1/2}}+2\frac{\gamma}{\alpha}\int_0^t\bar{A}_1(s)\bar{B}^2(s)\,ds.\end{aligned}$$

Then, $\bar{A}_1(t) \geq A_1(t)$ and $\bar{B}(t) \geq B(t)$, for all t for which $\bar{A}_1(t)$ and $\bar{B}(t)$ are finite. Moreover, $\bar{A}_1$ and $\bar{B}$ may be determined explicitly as,

$$\bar{A}_1(t) = \frac{\alpha}{1-\varepsilon^{1/2}e^{\gamma t}} \quad \text{and} \quad \bar{B}(t) = \frac{\beta e^{\gamma t}}{1-\varepsilon^{1/2}e^{\gamma t}},$$

whenever $\exp(\gamma t) < \varepsilon^{-1/2}$. Therefore, if M is chosen so that

$$M > 2\max\{\alpha, \beta e^{\gamma T}\},$$

and then ε_2 is chosen so that, as well as satisfying (4.26),

$$1-\varepsilon_2^{1/2}e^{\gamma T} \geq \frac{1}{2},$$

then $t_M = T$ for all ε in $(0, \varepsilon_2]$. Taking $c_4 = M$, the lemma is now established. □

The constants ε_2 and c_4 in Lemma 4.4 depend on $\|u_t(\cdot,0)\|_1$, since the constant a_4 in Lemma 4.3 had such a dependence. In order to control the size of $A(t)$ and $B(t)$, uniformly for small ε, some estimate of $\|u_t(\cdot,0)\|_1$ must be obtained in terms of the data f and g. An appropriate bound is forthcoming if the data satisfies the additional compatibility condition,

$$g_t(0) = -[f_x(0) + f(0)f_x(0) + f_{xxx}(0)]. \tag{4.28}$$

LEMMA 4.5. *Let $T>0$, $f \in H^\infty(\mathbb{R}^+)$, $g \in H^\infty(0,T)$ with $f(0)=g(0)$. Suppose the data f and g also satisfy* (4.28). *Then there is a constant a_5 depending on $\|f\|_4$ such that*

$$\|u_t(\cdot,0)\|_1 \leq a_5,$$

for all ε in $(0,1]$, *where u is the solution of* (4.1) *corresponding to f and g.*

Proof. Let $\varphi(x) = -[f_x(x) + f(x)f_x(x) + f_{xxx}(x)]$. Then $u_t(\cdot,0)$ is a solution of the boundary-value problem

$$u_t(\cdot,0) - \varepsilon u_{xxt}(\cdot,0) = \varphi,$$
$$u_t(0,0) = g_t(0), \qquad \lim_{x\to\infty} u_t(x,0) = 0.$$

Hence, $u_t(\cdot,0)$ is given by

$$u_t(x,0) = e^{-x/\varepsilon^{1/2}}g_t(0) + \int_0^\infty M_\varepsilon(x,\xi)\varphi(\xi)\,d\xi, \tag{4.29}$$

where, as in (3.10),

$$M_\varepsilon(x,\xi) = \frac{1}{2\varepsilon^{1/2}}\left[\exp\left(-|x-\xi|/\varepsilon^{1/2}\right) - \exp\left(-(x+\xi)/\varepsilon^{1/2}\right)\right].$$

It follows immediately from this representation that

$$\|u_t(\cdot,0)\| \leq \frac{\varepsilon^{1/4}}{2^{1/2}}|g_t(0)| + c\|\varphi\|,$$

where c is a constant which is independent of ε. Since $g_t(0) = \varphi(0)$, and because of the definition of φ, it is concluded there is a constant a depending on $\|f\|_4$ such that

$$\|u_t(\cdot,0)\| \leq a, \tag{4.30}$$

and this relation holds uniformly for ε in $(0,1]$. Differentiation of (4.29) with respect to x leads to the relation

$$u_{xt}(x,0) = -\frac{1}{\varepsilon^{1/2}} e^{-x/\varepsilon^{1/2}} g_t(0) + \frac{1}{2\varepsilon}\int_0^\infty e^{-(x+\xi)/\varepsilon^{1/2}}\varphi(\xi)\,d\xi$$
$$-\frac{1}{2\varepsilon}\int_0^x e^{(-x+\xi)/\varepsilon^{1/2}}\varphi(\xi)\,d\xi + \frac{1}{2\varepsilon}\int_x^\infty e^{(x-\xi)/\varepsilon^{1/2}}\varphi(\xi)\,d\xi.$$

Integrating the right-hand side by parts, there appears the formula

$$u_{xt}(x,0) = \frac{1}{\varepsilon^{1/2}} e^{-x/\varepsilon^{1/2}}\left[\varphi(0) - g_t(0)\right] + \int_0^\infty \tilde{M}_\varepsilon(x,\xi)\varphi_x(\xi)\,d\xi, \tag{4.31}$$

where

$$\tilde{M}_\varepsilon(x,\xi) = \frac{1}{2\varepsilon^{1/2}}\left[\exp\left(-|x-\xi|/\varepsilon^{1/2}\right) + \exp\left(-(x+\xi)/\varepsilon^{1/2}\right)\right].$$

The integral on the right-hand side of (4.31) presents no difficulty. For it is readily verified that

$$\left\| \int_0^\infty \tilde{M}_\varepsilon(\cdot,\xi)\varphi_x(\xi)\,d\xi \right\| \le c\|\varphi_x\|,$$

where again c denotes a constant independent of ε and φ. The presumption (4.28) has the effect of eliminating the other, potentially troublesome term from the right-hand side of (4.31). Again taking account of the definition of φ, it follows that there is a constant $\tilde{a}$, depending on $\|f\|_4$, such that

$$\|u_{xt}(\cdot,0)\| \le \tilde{a}, \tag{4.32}$$

holding uniformly for ε in $(0,1]$. Taken together, (4.30) and (4.32) imply the desired result. □

Combining the imports of Lemmas 4.4 and 4.5 leads directly to the principal result of this section.

THEOREM 4.6. *Let $T>0$ be given, and let $f \in H^\infty(\mathbb{R}^+)$ and $g \in H^\infty(0,T)$ and suppose the compatibility conditions*

$$f(0) = g(0), \qquad g_t(0) + f_x(0) + f(0)f_x(0) + f_{xxx}(0) = 0$$

hold. Let u be the solution of the regularized initial- and boundary-value problem (4.1) *corresponding to the given data f and g. Then there is a constant a_6, depending on $\|f\|_4$ and $|g|_{2,T}$, such that*

$$\|u(\cdot,t)\|_3 + \|u_t(\cdot,t)\|_1 \le a_6,$$

for all t in $[0,T]$ and ε in $(0,\varepsilon_2]$. Here ε_2 is the positive constant arising in Lemma 4.4, *and so depends on $\|f\|_4$ and $|g|_{2,T}$ as well.*

Remarks. A somewhat stronger result than is stated in Theorem 4.6 is available from the foregoing analysis. This strengthened result has been eschewed, for simplicity and because it is not needed in what follows. Nevertheless, it is worth recording that

$$\varepsilon\|u_{xxxx}(\cdot,t)\|^2 + \int_0^T \left[u_{xxx}^2(0,s) + u_{xxxx}^2(0,s) + \varepsilon u_{xxt}^2(0,s) + u_{xt}^2(0,s)\right] ds \le (a_6)^2$$

as well, provided that ε lies in $(0,\varepsilon_2]$ and t lies in $[0,T]$. The constants ε_2 and a_6 are those specified in the statement of the last theorem.

The various constants appearing in the statements of results in this section may all be taken to depend continuously and monotonically on both T and the norms of the data that occur. This follows immediately upon examination of the presented proofs. Such an aspect is without crucial significance in what follows, and so will be passed over.

5. Higher-order estimates for the regularized problem. The derivation of ε-independent bounds for solutions of the regularized initial- and boundary-value problem (4.1) is continued in this section. The bounds established in §4 would be sufficient to establish an existence theory set in the space $L^\infty(0,T;H^4(\mathbb{R}^+))$ for the quarter-plane problem (1.3). Smoother solutions would be expected to obtain provided the initial and boundary data is appropriately restricted. A proof of such further regularity, presented in §6, is based on the additional estimates to be obtained in the present section.

The assumption that $f\in H^\infty(\mathbb{R}^+)$, $g\in H^\infty(0,T)$, and $f(0)=g(0)$ will continue to be enforced throughout this section. This hypothesis will be recalled informally by the stipulation that the data f and g is smooth and compatible. If j is a nonnegative integer, the notation

$$u^{(j)}=\partial_t^j u$$

will be convenient, and employed henceforth. This section consists of two technical lemmas, which lead directly to the principal goal, Theorem 5.3. The first technical result generalizes Lemma 4.4.

LEMMA 5.1. *Let $f\in H^\infty(\mathbb{R}^+)$ and $g\in H^\infty(0,T)$ be given, with $f(0)=g(0)$. Let u be the solution of* (4.1) *corresponding to the data f and g, and let k be a nonnegative integer. There is a constant*

$$b_1=b_1\Big(|g|_{k+2,T},\ \max_{0\le j\le k}\big\{\|u^{(j)}(\cdot,0)\|_4,\|u^{(j+1)}(\cdot,0)\|_1\big\}\Big),$$

depending continuously on its arguments, such that

$$\|u^{(k)}(\cdot,t)\|_3^2+\varepsilon\|u^{(k)}_{xxxx}(\cdot,t)\|^2 + \int_0^t\Big\{\big[u^{(k)}_{xxx}(0,s)\big]^2+\big[u^{(k)}_{xxxx}(0,s)\big]^2+\varepsilon\big[u^{(k+1)}_{xx}(0,s)\big]^2\Big\}\,ds\le b_1,$$

$$\|u^{(k+1)}(\cdot,t)\|_1^2+\int_0^t\big[u_x^{(k+1)}(0,s)\big]^2 ds\le b_1,$$

for all t in $[0,T]$ and ε in $(0,\varepsilon_2]$. Here, ε_2 is specified in Lemma 4.4.

Proof. First note that for $k=0$, the desired result is implied by Lemma 4.4. The proof proceeds by induction on k. Let $k\ge 1$ be given, and suppose that the stated estimates hold for all nonnegative integers less than or equal to $k-1$. Let $v=u^{(k)}$, where u is the solution of the regularized initial- and boundary-value problem (4.1) corresponding to the given smooth and compatible data f and g. For t in $[0,T]$, define

$$A^2(t)=\sup_{0\le s\le t}\Big\{\|v(\cdot,s)\|_3^2+\varepsilon\|v_{xxxx}(\cdot,s)\|^2\Big\} + \int_0^t\big[v_{xxx}^2(0,s)+v_{xxxx}^2(0,s)+\varepsilon v_{xxt}^2(0,s)\big]\,ds$$

and

$$B^2(t)=\sup_{0\le s\le t}\Big\{\|v_t(\cdot,s)\|_1^2\Big\}+\int_0^t v_{xt}^2(0,s)\,ds.$$

The induction hypothesis implies that

$$\|u\|_{L^\infty(0,T;H^3(\mathbb{R}^+))},\ \|v\|_{L^\infty(0,T;H^1(\mathbb{R}^+))}\le c, \tag{5.1}$$

$$\|u\|_{L^\infty(0,T;W^{2,\infty}(\mathbb{R}^+))},\ \|v\|_{L^\infty(\mathbb{R}^+\times[0,T])}\le c,$$

where here, and in the remainder of this proof, c will denote various constants which all depend on the same variables as the constant b_1 given in the statement of the lemma, but which will always be independent of ε.

For any integer $j\ge 1$ the function $u^{(j)}$ satisfies the equation

$$u_t^{(j)}+u_x^{(j)}+\left(uu^{(j)}+h_j(u)\right)_x+u_{xxx}^{(j)}-\varepsilon u_{xxt}^{(j)}=0, \tag{5.2}$$

where

$$h_j(u)=\frac{1}{2}\sum_{i=1}^{j-1}\binom{j}{i}u^{(i)}u^{(j-i)}.$$

The induction hypothesis also implies that

$$\|h_k(u)\|_{L^\infty(0,T;W^{2,\infty}(\mathbb{R}^+))}\le c\|h_k(u)\|_{L^\infty(0,T;H^3(\mathbb{R}^+))}\le c. \tag{5.3}$$

The functions $A(t)$ and $B(t)$ will be estimated via an energy inequality derived from equation (5.2). Taking $j=k$, differentiate (5.2) once with respect to x, multiply by $-2v_{xxx}$ and integrate the resulting expression over $\mathbb{R}^+\times(0,t)$. The outcome of this process may be written

$$\begin{aligned} V_2(t)+\int_0^t\left[v_{xx}^2(0,s)+v_{xxx}^2(0,s)\right]ds \\ =V_2(0)-2\int_0^t v_{xt}(0,s)v_{xx}(0,s)\,ds+2\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xx}v_{xxx}\,dx\,ds, \end{aligned} \tag{5.4}$$

where $V_2(t)=\|v_{xx}(\cdot,t)\|^2+\varepsilon\|v_{xxx}(\cdot,t)\|^2$.

Inequalities (5.1) and (5.3) imply that

$$\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xx}v_{xxx}\,dx\,ds\le c\left(1+\int_0^t\|v(\cdot,s)\|_3^2 ds\right). \tag{5.5}$$

Because of (2.1) and (5.1), for any $\delta>0$, there is a constant c_δ such that for all t in $[0,T]$,

$$\|v\|_{L^\infty(0,t;W^{2,\infty}(\mathbb{R}^+))}\le c_\delta+\delta\left\{\sup_{0\le s\le t}\|v(\cdot,s)\|_3^2\right\}. \tag{5.6}$$

Combining (5.1), (5.2), (5.3), and (5.6), it follows that, for all $\delta>0$ and $t\in[0,T]$,

$$\begin{aligned} &-\int_0^t v_{xt}(0,s)v_{xx}(0,s)\,ds \\ &=\int_0^t\left\{v_{xx}(0,s)+\left[uv+h_k(u)\right]_{xx}(0,s)+v_{xxxx}(0,s)-\varepsilon v_{xxxt}(0,s)\right\}v_{xx}(0,s)\,ds \\ &\le c_\delta+\delta\left\{\sup_{0\le s\le t}\|v(\cdot,s)\|_3^2+\int_0^t v_{xxxx}^2(0,s)\,ds\right\}-\varepsilon\int_0^t v_{xxxt}(0,s)v_{xx}(0,s)\,ds. \end{aligned}$$

Together with (5.4) and (5.5) this implies that for all $\delta>0$ there is a constant c_δ such that

$$V_2(t)+\int_0^t\left[v_{xx}^2(0,s)+v_{xxx}^2(0,s)\right]ds \tag{5.7}$$
$$\leq c_\delta\left[1+\int_0^t A^2(s)\,ds\right]+\delta A^2(t)-2\varepsilon\int_0^t v_{xxxt}(0,s)v_{xx}(0,s)\,ds,$$

where A is defined above (5.1).

Next, differentiate (5.2), again with $j=k$, twice with respect to x, multiply by $-2v_{xxxx}$ and integrate over $\mathbb{R}^+\times(0,t)$. After suitable integrations by parts, there appears

$$V_3(t)+\int_0^t\left[v_{xxx}^2(0,s)+v_{xxxx}^2(0,s)\right]ds \tag{5.8}$$
$$=V_3(0)-2\int_0^t v_{xxt}(0,s)v_{xxx}(0,s)\,ds+2\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xxx}v_{xxxx}\,dx\,ds,$$

holding for all $t\in[0,T]$, and where

$$V_3(t)=\|v_{xxx}(\cdot,t)\|^2+\varepsilon\|v_{xxxx}(\cdot,t)\|^2.$$

Observe that

$$\int_0^t\int_0^\infty(uv)_{xxx}v_{xxxx}\,dx\,ds$$
$$=\int_0^t\int_0^\infty(uv_{xxx}+3u_xv_{xx}+3u_{xx}v_x+u_{xxx}v)v_{xxxx}\,dx\,ds$$
$$=-\int_0^t\Big[\frac{1}{2}g(s)v_{xxx}^2(0,s)+3u_x(0,s)v_{xx}(0,s)v_{xxx}(0,s)$$
$$+3u_{xx}(0,s)v_x(0,s)v_{xxx}(0,s)+u_{xxx}(0,s)v(0,s)v_{xxx}(0,s)\Big]ds$$
$$-\int_0^t\int_0^\infty\Big[\frac{7}{2}u_xv_{xxx}^2+6u_{xx}v_{xx}v_{xxx}+4u_{xxx}v_xv_{xxx}+u_{xxxx}vv_{xxx}\Big]dx\,ds.$$

The induction hypothesis and the fact that

$$\int_0^t\|v_x(\cdot,s)\|_{L^\infty(\mathbb{R}^+)}^2ds\leq c\int_0^t A^2(s)\,ds$$

implies that there is a constant c such that

$$\int_0^t\int_0^\infty u_{xxx}v_xv_{xxx}\,dx\,ds\leq\int_0^t\|v_x(\cdot,s)\|_{L^\infty(\mathbb{R}^+)}\|u_{xxx}(\cdot,s)\|\|v_{xxx}(\cdot,s)\|\,ds$$
$$\leq c\int_0^t A^2(s)\,ds.$$

Also, it follows directly from the regularized equation (4.1a) that

$$u_{xxxx}=\varepsilon u_{xxx}^{(1)}-\left(uu_{xx}+u_x^2+u_{xx}+u_x^{(1)}\right).$$

Hence, from (5.1) and the induction hypothesis,

$$\int_0^t\int_0^\infty u_{xxxx}vv_{xxx}\,dx\,ds\leq\int_0^t\|v(\cdot,s)\|_{L^\infty(\mathbb{R}^+)}\|u_{xxxx}(\cdot,s)\|\|v_{xxx}(\cdot,s)\|\,ds$$
$$\leq c\int_0^t A^2(s)\,ds,$$

for all t in $[0,T]$. By (5.1) and the above estimates, it may now be concluded that

$$\int_0^t\int_0^\infty (uv)_{xxx}v_{xxxx}\,dx\,ds \le c\Big[1+\int_0^t A^2(s)\,ds+\int_0^t v_{xxx}^2(0,s)\,ds\Big]. \tag{5.9}$$

To estimate the rest of the third term on the right-hand side of (5.8), note that

$$\int_0^t\int_0^\infty (h_k(u))_{xxx}v_{xxxx}\,dx\,ds$$
$$= -\int_0^t (h_k(u))_{xxx}(0,s)v_{xxx}(0,s)\,ds-\int_0^t\int_0^\infty (h_k(u))_{xxxx}v_{xxx}\,dx\,ds.$$

Equation (5.2), once-differentiated with respect to x, is

$$u_{xxxx}^{(j)}=\varepsilon u_{xxx}^{(j+1)}-\big\{\big[uu^{(j)}+h_j(u)\big]_{xx}+u_{xx}^{(j)}+u_x^{(j+1)}\big\}.$$

Together with the induction hypothesis this relation implies that

$$\int_0^t \|(h_k(u))_{xxxx}(\cdot,s)\|^2\,ds \le c\Big[1+\varepsilon^2\int_0^t A^2(s)\,ds\Big].$$

Therefore, using again the induction hypothesis and the estimate above, we may conclude that

$$\int_0^t\int_0^\infty (h_k(u))_{xxx}v_{xxxx}\,dx\,ds \le c\Big[1+\int_0^t A^2(s)\,ds+\int_0^t v_{xxx}^2(0,s)\,ds\Big]. \tag{5.10}$$

It remains to estimate the boundary term on the right-hand side of (5.8). The equation (5.2), with $j=k$ again, implies

$$-\int_0^t v_{xxt}(0,s)v_{xxx}(0,s)\,ds$$
$$=\int_0^t v_{xxt}(0,s)\{v_t(0,s)+v_x(0,s)+[uv+h_k(u)]_x(0,s)-\varepsilon v_{xxt}(0,s)\}\,ds.$$

Integrating by parts with respect to s yields the relation

$$\int_0^t v_{xxt}(0,s)\{v_t(0,s)+v_x(0,s)+[uv+h_k(u)]_x(0,s)\}\,ds$$
$$=v_{xx}(0,s)\{v_t(0,s)+v_x(0,s)+[uv+h_k(u)]_x(0,s)\}\Big|_{s=0}^{s=t}$$
$$-\int_0^t v_{xx}(0,s)\{v_{tt}(0,s)+v_{xt}(0,s)+[uv+h_k(u)]_{xt}(0,s)\}\,ds.$$

From (5.1), (5.3) and (5.6), and the fact that $v_{tt}(0,s)=g^{(k+2)}(s)$ and $v_t(0,s)=g^{(k+1)}(s)$, it thus appears that for any $\delta>0$ there is a constant c_δ such that

$$-\int_0^t v_{xxt}(0,s)v_{xxx}(0,s)\,ds$$
$$\le c_\delta-\varepsilon\int_0^t v_{xxt}^2(0,s)\,ds+\delta A^2(t)-\int_0^t[1+g(s)]v_{xx}(0,s)v_{xt}(0,s)\,ds. \tag{5.11}$$

The estimates (5.8), (5.9), (5.10) and (5.11) and the identity

$$-v_{xt}=v_{xx}+[uv+h_k(u)]_{xx}+v_{xxxx}-\varepsilon v_{xxxt},$$

obtained from (5.2), now imply that, for all $\delta>0$, there is a constant c_δ such that for all $t\in[0,T]$,

$$V_3(t)+\int_0^t\left[v_{xxx}^2(0,s)+v_{xxxx}^2(0,s)+\varepsilon v_{xxt}^2(0,s)\right]ds$$
$$\leq c_\delta\left[1+\int_0^t A^2(s)\,ds+\int_0^t v_{xxx}^2(0,s)\,ds\right]+\delta A^2(t)$$
$$-2\varepsilon\int_0^t[1+g(s)]\,v_{xx}(0,s)\,v_{xxxt}(0,s)\,ds.$$

By adding this estimate and a suitable multiple of (5.7), and using the induction hypothesis again, it appears that for each $\delta>0$ there is a constant c_δ so that, for all t in $[0,T]$,

$$A^2(t)\leq c_\delta\left[1+\int_0^t A^2(s)\,ds\right]+\delta\varepsilon^2\int_0^t v_{xxxt}^2(0,s)\,ds. \tag{5.12}$$

Inequality (5.12) is not useful until the second integral is bounded. This may be accomplished by virtually the same argument as was used to bound the corresponding term appearing in the proof of Lemma 4.2. Differentiate (5.2), with $j=k$, twice with respect to x, multiply the result by $2\varepsilon v_{xxxt}$, and then integrate over $\mathbb{R}^+\times(0,t)$. This leads to the identity

$$\varepsilon\left\{\|v_{xxx}(\cdot,t)\|^2-\|v_{xxxx}(\cdot,t)\|^2\right\}+\varepsilon^2\int_0^t v_{xxxt}^2(0,s)\,ds \tag{5.13}$$
$$=\varepsilon\left\{\|v_{xxx}(\cdot,0)\|^2-\|v_{xxxx}(\cdot,0)\|^2\right\}+\varepsilon\int_0^t v_{xxt}^2(0,s)\,ds$$
$$+2\varepsilon\int_0^t v_{xxxx}(0,s)\,v_{xxxt}(0,s)\,ds-2\varepsilon\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xxx}v_{xxxt}\,dx\,ds.$$

Since (5.2) implies that

$$\varepsilon\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xxx}v_{xxxt}\,dx\,ds$$
$$=\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xxx}\left\{v_{xxxx}+\left[uv+h_k(u)\right]_{xx}+v_{xx}+v_{xt}\right\}dx\,ds,$$

it follows from (5.9), (5.10) and the induction hypothesis that for all $t\in[0,T]$,

$$\varepsilon\int_0^t\int_0^\infty\left[uv+h_k(u)\right]_{xxx}v_{xxxt}\,dx\,ds$$
$$\leq c\left\{1+\int_0^t\left[A^2(s)+A(s)B(s)\right]ds+\int_0^t v_{xxx}^2(0,s)\,ds\right\}.$$

In consequence of (5.12) and (5.13) we therefore infer the existence of a constant c such that

$$A^2(t)\leq c\left\{1+\int_0^t\left[A^2(s)+A(s)B(s)\right]ds\right\}, \tag{5.14}$$

for all $t\in[0,T]$.

Next $B(t)$ will be estimated. Let $w=u^{(k+1)}$. By (5.2) w satisfies the equation

$$w_t+w_x+\left[uw+h_{k+1}(u)\right]_x+w_{xxx}-\varepsilon w_{xxt}=0. \tag{5.15}$$

Multiply this equation by $2w$ and integrate over $\mathbb{R}^+\times(0,t)$ to obtain

$$\begin{aligned}\|w(\cdot,t)\|^2+\varepsilon\|w_x(\cdot,t)\|^2+\int_0^t w_x^2(0,s)\,ds&\\ =\|w(\cdot,0)\|^2+\varepsilon\|w_x(\cdot,0)\|^2+\int_0^t[1+g(s)]\,w^2(0,s)\,ds&\\ +2\int_0^t w(0,s)\big[w_{xx}(0,s)-\varepsilon w_{xt}(0,s)\big]\,ds&\\ -\int_0^t\int_0^\infty\{u_xw^2+2w[h_{k+1}(u)]_x\}\,dx\,ds.&\end{aligned}$$

The induction hypothesis therefore implies that, for all $\delta>0$, there is a constant c_δ such that

$$\begin{aligned}(5.16)\quad \|w(\cdot,t)\|^2+\varepsilon\|w_x(\cdot,t)\|^2+\int_0^t w_x^2(0,s)\,ds&\\ \le c_\delta\Big[1+\int_0^t B^2(s)\,ds\Big]+\delta\int_0^t\big[w_{xx}(0,s)-\varepsilon w_{xt}(0,s)\big]^2ds,&\end{aligned}$$

for all $t\in[0,T]$. To complete the satisfactory estimation of $B(t)$, multiply (5.15) by $2(\varepsilon w_{xt}-uw-w_{xx})$ and integrate over $\mathbb{R}^+\times(0,t)$. This yields

$$\begin{aligned}(5.17)\quad (1+\varepsilon)\|w_x(\cdot,t)\|^2-\int_0^\infty w^2(x,t)u(x,t)\,dx&\\ +\int_0^t\{w_x^2(0,s)+[w_{xx}(0,s)-\varepsilon w_{xt}(0,s)]^2\}\,ds&\\ =(1+\varepsilon)\|w_x(\cdot,0)\|^2-\int_0^\infty w^2(x,0)f(x)\,dx&\\ +\int_0^t[\varepsilon w_t^2(0,s)-g^2(s)w^2(0,s)]\,ds&\\ -2\int_0^t w_t(0,s)w_x(0,s)\,ds-2\int_0^t g(s)w(0,s)[w_{xx}(0,s)-\varepsilon w_{xt}(0,s)]\,ds&\\ +\int_0^t\int_0^\infty\{2uww_x-u_tw^2+2[h_{k+1}(u)]_x(w_{xx}+uw-\varepsilon w_{xt})\}\,dx\,ds.&\end{aligned}$$

Integration by parts implies that

$$\begin{aligned}\int_0^t\int_0^\infty[h_{k+1}(u)]_xw_{xx}\,dx&\\ =-\int_0^t[h_{k+1}(u)]_x(0,s)w_x(0,s)\,ds-\int_0^t\int_0^\infty[h_{k+1}(u)]_{xx}w_x\,dx\,ds,&\end{aligned}$$

and that

$$\begin{aligned}\varepsilon\int_0^t\int_0^\infty[h_{k+1}(u)]_xw_{xt}\,dx\,ds=\varepsilon\int_0^\infty[h_{k+1}(u)]_x(x,s)w_x(x,s)\,dx\Big|_{s=0}^{s=t}&\\ -\varepsilon\int_0^t\int_0^\infty[h_{k+1}(u)]_{xt}w_x\,dx\,ds.&\end{aligned}$$

Hence, it follows from the induction hypothesis that for all $t\in[0,T]$,

$$\int_0^t\int_0^\infty [h_{k+1}(u)]_x(w_{xx}+uw-\varepsilon w_{xt})\,dx\,ds$$

$$\leq \varepsilon^2\|w_x(\cdot,t)\|^2+c\left\{1+\int_0^t\left[B^2(s)+A(s)B(s)\right]ds+\int_0^t w_x^2(0,s)\,ds\right\}.$$

Therefore, if (5.17) is added to a suitable multiple of (5.16), it follows that

$$B^2(t)\leq c\left\{1+\int_0^t\left[B^2(s)+A(s)B(s)\right]ds\right\} \tag{5.18}$$

for all $t\in[0,T]$ and all ε in $(0,\varepsilon_2]$. Here, without loss of generality, ε_2 has been presumed to be strictly less than 1.

From (5.14), (5.18) and Gronwall's lemma it now follows that there is a constant c such that

$$A(t),\,B(t)\leq c$$

for all $t\in[0,T]$. This completes the induction argument and hence the proof of Lemma 5.1. □

The bounds established in Lemma 5.1 are just what will be needed in §6, except that, so far as is known now, not all the arguments of the constant b_1 are independent of ε. To attain the goal for this section, it will suffice to give conditions on the data f and g which imply that $\|u^{(j)}(\cdot,0)\|_4$ and $\|u^{(j+1)}(\cdot,0)\|_1$, $0\leq j\leq k$, are bounded, independently of ε sufficiently small. This amounts to extending Lemma 4.5.

We have not succeeded in giving an absolutely straightforward generalization of Lemma 4.5 to the case $j>0$. However, by modifying the data, in an ε-dependent way, a result is obtained which is sufficient for our purposes in the next section. Before stating this lemma, some convenient notation is introduced.

Let $\varphi^{(0)}(x)=f(x)$, and for each integer $j\geq 1$ define functions $\varphi^{(j)}$ inductively by the recurrence

$$\varphi^{(j+1)}=-\left[\varphi_x^{(j)}+\varphi_{xxx}^{(j)}+\frac{1}{2}\left(\sum_{i=0}^{j}\binom{j}{i}\varphi^{(i)}\varphi^{(j-i)}\right)_x\right]. \tag{5.19}$$

Also, for nonnegative integers j, let

$$g^{(j)}(t)=\partial_t^j g(t).$$

Here is the result alluded to above.

LEMMA 5.2. *Let $f\in H^\infty(\mathbb{R}^+)$ and $g\in H^\infty(0,T)$ be given, with $f(0)=g(0)$. Let $k\geq 1$ be a given integer and suppose additionally that*

$$g^{(j)}(0)=\varphi^{(j)}(0)\quad \text{for } j=1,2,\cdots,k.$$

Then there exists a family $\{g_\varepsilon\}_{0<\varepsilon\leq 1}$ in $H^\infty(0,T)$ such that

(i) *$g_\varepsilon(0)=g(0)$ and $\lim_{\varepsilon\to 0}|g_\varepsilon-g|_{k+1,T}=0$;*

(ii) *there exists a constant b_2, depending continuously on $\|f\|_{3k+1}$, such that*

$$\|u_\varepsilon^{(j)}(\cdot,0)\|_{3(k-j)+1}\leq b_2$$

for $0\leq j\leq k$ and all $\varepsilon\in(0,1]$, where u_ε denotes the solution of (4.1) with initial data f and boundary data g_ε.

Proof. First, two sequences of functions $\{\varphi_\varepsilon^{(j)}\}_{1\le j\le k}$ and $\{w_\varepsilon^{(j)}\}_{1\le j\le k}$ are introduced. These will be used momentarily to define the modified boundary data $g_\varepsilon(t)$. If j is an integer in the range $[0,k]$, let $\nu(j)=[3(k-j)/2]$ and define $w_\varepsilon^{(j)}$ and $\varphi_\varepsilon^{(j)}$ on $\mathbb{R}^+$ by $w_\varepsilon^{(0)}=\varphi_\varepsilon^{(0)}=f$ and, recursively for $j>0$,

$$\varphi_\varepsilon^{(j)}=-\left[\left(w_\varepsilon^{(j-1)}\right)_x+\left(w_\varepsilon^{(j-1)}\right)_{xxx}+\frac{1}{2}\sum_{i=0}^{j-1}\binom{j-1}{i}\left(w_\varepsilon^{(i)}w_\varepsilon^{(j-i-1)}\right)_x\right] \tag{5.20}$$

and

$$w_\varepsilon^{(j)}=\exp\left(-x/\varepsilon^{1/2}\right)\sum_{i=0}^{\nu(j)}\varepsilon^i\left(\partial_x^{2i}\varphi_\varepsilon^{(j)}\right)(0)+\int_0^\infty M_\varepsilon(x,\xi)\varphi_\varepsilon^{(j)}(\xi)\,d\xi. \tag{5.21}$$

Here, as in the proof of Lemma 4.5,

$$M_\varepsilon(x,\xi)=\frac{1}{2\varepsilon^{1/2}}\left[\exp\left(-|x-\xi|/\varepsilon^{1/2}\right)-\exp\left(-(x+\xi)/\varepsilon^{1/2}\right)\right]$$

and

$$\tilde{M}_\varepsilon(x,\xi)=\frac{1}{2\varepsilon^{1/2}}\left[\exp\left(-|x-\xi|/\varepsilon^{1/2}\right)+\exp\left(-(x+\xi)/\varepsilon^{1/2}\right)\right].$$

Note that $w_\varepsilon^{(j)}$ has been determined as the solution of the boundary-value problem

$$v-\varepsilon v_{xx}=\varphi_\varepsilon^{(j)}, \tag{5.22}$$

with

$$v(0)=\lambda_\varepsilon^{(j)} \quad \text{and} \quad \lim_{x\to+\infty} v(x)=0,$$

where

$$\lambda_\varepsilon^{(j)}=\sum_{i=0}^{\nu(j)}\varepsilon^i\left(\partial_x^{2i}\varphi_\varepsilon^{(j)}\right)(0),$$

for $j=1,2,\cdots,k$.

By differentiating (5.21) the following identities are obtained, for all integers $r\ge 1$,

$$\left(\partial_x^{2r+1}w_\varepsilon^{(j)}\right)(x)=\exp\left(-x/\varepsilon^{1/2}\right)\varepsilon^{-(r+1/2)}\left[\sum_{i=0}^{r}\varepsilon^i\left(\partial_x^{2i}\varphi_\varepsilon^{(j)}\right)(0)-\lambda_\varepsilon^{(j)}\right]+\int_0^\infty\tilde{M}_\varepsilon(x,\xi)\left(\partial_x^{2r+1}\varphi_\varepsilon^{(j)}\right)(\xi)\,d\xi \tag{5.23a}$$

and

$$\left(\partial_x^{2r}w_\varepsilon^{(j)}\right)(x)=\exp\left(-x/\varepsilon^{1/2}\right)\varepsilon^{-r}\left[\lambda_\varepsilon^{(j)}-\sum_{i=0}^{r-1}\varepsilon^i\left(\partial_x^{2i}\varphi_\varepsilon^{(j)}\right)(0)\right]+\int_0^\infty M_\varepsilon(x,\xi)\left(\partial_x^{2r}\varphi_\varepsilon^{(j)}\right)(\xi)\,d\xi. \tag{5.23b}$$

Hence, there is a constant c, independent of $w_\varepsilon^{(j)}$, $\varphi_\varepsilon^{(j)}$ and ε, such that

$$\left\|w_\varepsilon^{(j)}\right\|_{3(k-j)+1}\le c\left\|\varphi_\varepsilon^{(j)}\right\|_{3(k-j)+1}, \tag{5.24}$$

for $0\le j\le k$. Using (5.20), (5.24) and a simple inductive argument, it follows that there is a constant $b_2=b_2(\|f\|_{3k+1})$ such that

$$\|w_\varepsilon^{(j)}\|_{3(k-j)+1},\ \|\varphi_\varepsilon^{(j)}\|_{3(k-j)+1}\le b_2, \tag{5.25}$$

independently of ε in $(0,1]$ and j in $[0,k]$.

For each $\varepsilon\in(0,1]$ define modified boundary data $g_\varepsilon(t)$ by

$$g_\varepsilon(t)=g(t)+\sum_{j=1}^{k}\frac{t^j}{j!}\left[\lambda_\varepsilon^{(j)}-\varphi^{(j)}(0)\right].$$

Observe that $g_\varepsilon(0)=g(0)$. Also, since $g^{(j)}(0)=\varphi^{(j)}(0)$ by assumption,

$$g_\varepsilon^{(j)}(0)=\lambda_\varepsilon^{(j)}, \tag{5.26}$$

for $1\le j\le k$.

Now let u_ε denote the solution of (4.1) with initial data f and boundary data g_ε. It follows inductively from (5.20), (5.22) and (5.26) that $u_\varepsilon^{(j)}(\ ,0)=w_\varepsilon^{(j)}$ for $0\le j\le k$, and hence the desired bounds on $u_\varepsilon^{(j)}(\cdot,0)$ follow from (5.25).

To complete the proof it is only required to check that

$$\lim_{\varepsilon\downarrow 0}|g_\varepsilon-g|_{k+1,T}=0.$$

Because of the definition of g_ε, this is equivalent to showing that

$$\lim_{\varepsilon\downarrow 0}|\lambda_\varepsilon^{(j)}-\varphi^{(j)}(0)|=0,$$

for $0\le j\le k$. Referring to the definition of $\lambda_\varepsilon^{(j)}$ below (5.22), and keeping in mind the bounds in (5.25) and the simple inequality (2.5), we see that

$$\lambda_\varepsilon^{(j)}=\varphi_\varepsilon^{(j)}(0)+O(\varepsilon),$$

as $\varepsilon\downarrow 0$, for $0\le j\le k$. More precisely,

$$|\lambda_\varepsilon^{(j)}-\varphi_\varepsilon^{(j)}(0)|\le c\varepsilon\|\varphi_\varepsilon^{(j)}\|_{3(k-j)+1}\le cb_2\varepsilon. \tag{5.27}$$

Hence it is enough to show that

$$\lim_{\varepsilon\downarrow 0}|\varphi_\varepsilon^{(j)}(0)-\varphi^{(j)}(0)|=0,$$

for $0\le j\le k$. This latter relation will be proved by establishing that the estimate

$$\|\varphi_\varepsilon^{(i)}-\varphi^{(i)}\|_{W^{3(k-i),\infty}(\mathbb{R}^+)}\le c\varepsilon^{1/4}, \tag{5.28}$$

holds for $0\le i\le k$, where the constant $c=c(\|f\|_{3k+1})$.

The inequality (5.28) is proved by induction on i. For $i=0$ and $i=1$, (5.28) follows since $\varphi_\varepsilon^{(0)}=\varphi^{(0)}=f$ and $\varphi_\varepsilon^{(1)}=\varphi^{(1)}$. Assume (5.28) holds for $i\le j$, where $1\le j<k$. In order to establish the result for $i=j+1$, note first that the definitions (5.19) and (5.20) imply that

$$\|\varphi_\varepsilon^{(j+1)}-\varphi^{(j+1)}\|_{W^{3(k-j-1),\infty}(\mathbb{R}^+)}\le c\left\{\sup_{0\le i\le j}\|w_\varepsilon^{(i)}-\varphi^{(i)}\|_{W^{3(k-i),\infty}(\mathbb{R}^+)}\right\},$$

where $c=c(\|f\|_{3k+1})$. Since

$$\|\varphi_\varepsilon^{(i)}-\varphi^{(i)}\|_{W^{3(k-i),\infty}(\mathbb{R}^+)}\le c\varepsilon^{1/4},$$

for $0\le i\le j$, by the induction hypothesis, (5.28) will follow if it can be demonstrated that, for $0\le i\le j$,

$$\|\varphi_\varepsilon^{(i)}-w_\varepsilon^{(i)}\|_{W^{3(k-i),\infty}(\mathbb{R}^+)}\le c\varepsilon^{1/4}, \tag{5.29}$$

where again $c=c(\|f\|_{k+1})$. The fact that $w_\varepsilon^{(i)}$ solves (5.22) means that

$$w_\varepsilon^{(i)}(x)-\varphi_\varepsilon^{(i)}(x)=\exp(-x/\varepsilon^{1/2})\big[\lambda_\varepsilon^{(i)}-\varphi_\varepsilon^{(i)}(0)\big]+\varepsilon\int_0^\infty M_\varepsilon(x,\xi)\partial_x^2\varphi_\varepsilon^{(i)}(\xi)\,d\xi.$$

Differentiating this relation with respect to x, in the same way that (5.21) was differentiated to yield (5.23a, b), and using (5.27), we readily obtain the estimate,

$$\|w_\varepsilon^{(i)}-\varphi_\varepsilon^{(i)}\|_{3(k-i)-1}\le c\varepsilon\|\varphi_\varepsilon^{(i)}\|_{3(k-i)+1},$$

where the constant c is independent of $w_\varepsilon^{(i)}$, $\varphi_\varepsilon^{(i)}$ and ε. The bounds expressed in (5.25) thus imply that

$$\|w_\varepsilon^{(i)}-\varphi_\varepsilon^{(i)}\|_{3(k-i)-1}\le c\varepsilon, \tag{5.30}$$

where $c=c(\|f\|_{3k+1})$. Also implied by (5.25), and the triangle inequality, is the estimate

$$\|w_\varepsilon^{(i)}-\varphi_\varepsilon^{(i)}\|_{3(k-i)+1}\le c, \tag{5.31}$$

where $c=c(\|f\|_{3k+1})$. Standard results in the interpolation-theory of Banach spaces now come to our rescue (cf. (2.5) and [19, Chap. 1]). Thus, if h denotes $\varphi_\varepsilon^{(i)}-w_\varepsilon^{(i)}$, then

$$\begin{aligned}\|h\|_{W^{3(k-i),\infty}(\mathbb{R}^+)}&\le\|h\|_{3(k-i)}^{1/2}\|h\|_{3(k-i)+1}^{1/2}\\&\le c\|h\|_{3(k-i)-1}^{1/4}\|h\|_{3(k-i)+1}^{3/4}\le c\varepsilon^{1/4},\end{aligned}$$

where $c=c(\|f\|_{3k+1})$. This completes the induction argument in favor of (5.28), and thus finishes the proof of the lemma. □

The outcome of Lemmas 5.1 and 5.2 is conveniently collected in the following theorem. This is, in effect, a higher-order analogue of Theorem 4.6. In the statement of the theorem, ε_2 is the same positive constant that already appeared in Theorem 4.6.

THEOREM 5.3. *Let $T>0$ and a positive integer k be given. Let $f\in H^\infty(\mathbb{R}^+)$ and $g\in H^\infty(0,T)$ and suppose that $g^{(j)}(0)=\varphi^{(j)}(0)$, for $0\le j\le k$, where the functions $\varphi^{(j)}$ are related to f as in* (5.19). *Then there exists a family $\{g_\varepsilon\}_{0<\varepsilon\le\varepsilon_2}$ in $H^\infty(0,T)$ such that*

(i) $g_\varepsilon(0)=g(0)$, $\lim_{\varepsilon\downarrow 0}|g_\varepsilon-g|_{k+1,T}=0$;

(ii) *there exists a constant $b_3=b_3(\|f\|_{3k+1},|g|_{k+1,T})$, depending continuously on its arguments, such that*

$$\begin{aligned}&\|u^{(j-1)}(\cdot,t)\|_3^2+\varepsilon\|\partial_x^4u^{(j-1)}(\cdot,t)\|^2+\|u^{(j)}(\cdot,t)\|_1^2\\&\quad+\int_0^t\Big\{\big[\partial_x^4u^{(j-1)}(0,s)\big]^2+\big[\partial_x^3u^{(j-1)}(0,s)\big]^2+\big[\partial_xu^{(j)}(0,s)\big]^2\\&\qquad+\varepsilon\big[\partial_x^2u^{(j)}(0,s)\big]^2\Big\}\,ds\le b_3\end{aligned}$$

holds for $1\le j\le k$ and all ε in $(0,\varepsilon_2]$. Here, $u^{(j-1)}(x,t)=\partial_t^{j-1}u_\varepsilon(x,t)$ and u_ε denotes the solution of (4.1) *with initial data f and boundary data g_ε.*

6. Existence and uniqueness of solution. The major undertaking of this paper is to prove existence of smooth solutions of the quarter-plane problem for the KdV equation. Using the theory developed in §§3, 4 and 5, this task becomes comparatively

simple. Recall that a function $u=u(x,t)$ is sought such that

$$u_t+u_x+uu_x+u_{xxx}=0 \quad \text{for } x,t>0, \tag{6.1a}$$

subject to the auxiliary conditions,

$$\begin{aligned} u(x,0)&=f(x) \quad \text{for } x\geq 0,\\ u(0,t)&=g(t) \quad \text{for } t\geq 0, \end{aligned} \tag{6.1b}$$

where f and g are given functions.

The issue of uniqueness of solutions of this initial- and boundary-value problem is especially straightforward to settle. As the uniqueness of solutions of (6.1) is useful later, it is established first.

THEOREM 6.1. *Let $T>0$ and $s>\frac{3}{2}$. Then, corresponding to given auxiliary data f and g, there is at most one solution of* (6.1) *in the function class $L^\infty(0,T;H^s(\mathbb{R}^+))$.*

Remarks. As usual in this paper, we mean, at the outset, by the word *solution* a distributional solution of (6.1a) for which the auxiliary conditions (6.1b) can be given a well-defined sense. Of course if u is a distributional solution of (6.1a) which is additionally known to lie in a class of smooth functions, it will follow that u is a classical solution of the differential equation. This point will be amplified later in this section.

Proof. Suppose that $u,v\in L^\infty(0,T;H^s(\mathbb{R}^+))$ are both solutions of (6.1) corresponding to the same data f and g. The $H^s(\mathbb{R}^+)$-norm of u and v is thus essentially bounded on $[0,T]$. In particular, for almost every t in $[0,T]$, $u(\cdot,t)$, $v(\cdot,t)\in H^s(\mathbb{R}^+)$. Invoking the Sobolev embedding results (cf. [19, Chap. 1]), it may therefore be supposed that, for almost every t in $[0,T]$, $u(\cdot,t)$, $u_x(\cdot,t)$, $v(\cdot,t)$ and $v_x(\cdot,t)$ are bounded and uniformly continuous functions on $\overline{\mathbb{R}}^+$. Moreover, u,u_x, v and v_x are essentially bounded on $\overline{\mathbb{R}}^+\times[0,T]$. From this it follows straightforwardly that both u and v converge, in $L^\infty(0,T)$, in the limit as $x\downarrow 0$. Thus the boundary value in (6.1b) is taken on meaningfully.

Let $w=u-v$ and $\chi=\frac{1}{2}(u+v)$. Then w is a distributional solution of the linear variable-coefficient differential equation

$$w_t+w_x+(\chi w)_x+w_{xxx}=0 \quad \text{in } \mathbb{R}^+\times(0,T), \tag{6.2a}$$

which satisfies the auxiliary conditions

$$w(x,0)=0 \quad \text{for } x\in\mathbb{R}^+, \qquad w(0,t)=0 \quad \text{for } t \text{ in } [0,T]. \tag{6.2b}$$

The boundary condition in (6.2b) holds at least in $L^\infty(0,T)$, whereas it will appear presently that the initial condition is valid at least in the sense that $\|w(\cdot,t)\|\to 0$, as $t\downarrow 0$.

Since $H^q(\mathbb{R}^+)$ is linearly and continuously embedded in $H^s(\mathbb{R}^+)$, for $q>s$, we may, without loss of generality, suppose that $s<3$ and let $r=3-s$. Note that $0<r<3/2$. Note also that w_x and $(\chi w)_x$ lie in $L^\infty(0,T;H^{s-1}(\mathbb{R}^+))$ and that w_{xxx} lies in $L^\infty(0,T;H^{-r}(\mathbb{R}^+))$. From (6.2a) it is thus apparent that w_t lies in $L^\infty(0,T;H^{-r}(\mathbb{R}^+))$.

The spaces $H_0^r(\mathbb{R}^+)$ and $H^{-r}(\mathbb{R}^+)$ are viewed as being in duality in the usual manner. The pairing between them is denoted by sharp brackets $\langle\cdot,\cdot\rangle$. (For a detailed exposition of these spaces, and the duality between them, the reader is urged to consult the first two chapters of Lions and Magenes [19].) Note especially that since, for almost every t in $[0,T]$, $w\in H^s(\mathbb{R}^+)$ and $w(0,t)=0$, it follows that $w\in H_0^r(\mathbb{R}^+)$, for almost every t in $[0,T]$. Thus $w\in L^\infty(0,T;H^s(\mathbb{R}^+)\cap H_0^r(\mathbb{R}^+))$. For this, it is crucial that $r<3/2$ of course. Otherwise a second boundary condition $w_x(0,t)=0$ would be implied by membership in $H_0^r(\mathbb{R}^+)$.

In this situation, it is a standard result (cf. [18, p. 71]) that $w \in C(0, T; L^2(\mathbb{R}^+))$, and that

$$\frac{1}{2}\frac{d}{dt}\|w(\cdot,t)\|^2 = \langle w, w_t \rangle. \tag{6.3}$$

Thus, in particular, the initial value in (6.1b) or (6.2b) is taken on meaningfully. The right-hand side of (6.3) lies in $L^1(0,T)$. Hence $\|w(\cdot,t)\|^2$ is absolutely continuous, and upon integrating (6.3) over $[0,t]$, using the equation (6.2a) and the zero initial condition in (6.2b), there appears

$$\frac{1}{2}\|w(\cdot,t)\|^2 = -\int_0^t \langle w, w_x + (\chi w)_x + w_{xxx} \rangle \, d\tau. \tag{6.4}$$

Since w_x and $(\chi w)_x$ are continuous square-integrable functions, for almost every t, and $w(0,t)=0$, it is straightforward that

$$\langle w, w_x \rangle = \int_0^\infty w(x,t) w_x(x,t)\, dx = 0,$$

and that

$$\begin{aligned}\langle w, (\chi w)_x \rangle &= \int_0^\infty w(x,t)[\chi(x,t) w(x,t)]_x \, dx \\ &= \frac{1}{2}\int_0^\infty w^2(x,t)\chi_x(x,t)\, dx \\ &\le \|\chi_x\|_{L^\infty(\mathbb{R}^+\times(0,T))}\|w(\cdot,t)\|^2 \le M\|w(\cdot,t)\|^2,\end{aligned}$$

where

$$M = \frac{1}{2}\|u+v\|_{L^\infty(0,T;H^s(\mathbb{R}^+))}.$$

In the last step, the fact that $s > \frac{3}{2}$ was vital. Finally, we claim that $\langle w, w_{xxx} \rangle \ge 0$, for almost every t in $[0,T]$. Fix t and let $h(\cdot) = w(\cdot,t)$. Then $h \in H^s(\mathbb{R}^+) \cap H_0^r(\mathbb{R}^+)$. Let $\tilde{h}$ be a function in $H^\infty(\mathbb{R}^+)$, say, such that

$$\partial_x^j \tilde{h}(0) = \partial_x^j h(0) \quad \text{for } 0 \le j < s - \frac{1}{2}.$$

Then $h - \tilde{h} \in H_0^s(\mathbb{R}^+)$. Hence there is a sequence $\{\psi_n\}_1^\infty$ in $\mathcal{D}(\mathbb{R}^+)$ such that $\psi_n \to h - \tilde{h}$ in the $H^s(\mathbb{R}^+)$-norm, as $n \to \infty$. Let $h_n = \psi_n + \tilde{h}$. The sequence $\{h_n\}_1^\infty$ has the following properties:

(i) $h_n \in H^\infty(\mathbb{R}^+)$ and $h_n(0) = 0$, for all n;

(ii) $h_n \to h$ in $H^s(\mathbb{R}^+)$, as $n \to \infty$.

Then $\partial_x^3 h_n \to \partial_x^3 h$ in $H^{-r}(\mathbb{R}^+)$ and $h_n \to h$ in $H_0^r(\mathbb{R}^+)$, as $n \to \infty$. Hence,

$$\begin{aligned}\langle h, h_{xxx} \rangle = \lim_{n\to\infty} \langle h_n, \partial_x^3 h_n \rangle &= \lim_{n\to\infty} \int_0^\infty h_n(x)\partial_x^3 h_n(x)\, dx \\ &= \lim_{n\to\infty} \left\{ -\int_0^\infty \partial_x h_n(x)\partial_x^2 h_n(x)\, dx \right\} \\ &= \lim_{n\to\infty} \frac{1}{2}[\partial_x h_n(0)]^2 \ge 0.\end{aligned}$$

Putting together the pieces, there appears

$$\|w(\cdot,t)\|^2 \le M\int_0^t \|w(\cdot,\tau)\|^2 d\tau,$$

for t in $[0,T]$. Gronwall's lemma thus implies that $\|w(\cdot,t)\| \equiv 0$ on $[0,T]$, whence $w=0$ and so $u=v$, as required. □

Attention is now turned to the existence theory. It is convenient to recall here the notation introduced in §5. Namely, if f is a given sufficiently smooth function defined on $\overline{\mathbb{R}}^+$, then set $\varphi^{(0)}=f$,

$$\varphi^{(1)}(x) = -\left(f(x) + \frac{1}{2} f^2(x) + f_{xx}(x) \right)_x, \tag{6.5a}$$

and inductively,

$$\varphi^{(j+1)}(x) = -\left(\varphi_x^{(j)} + \varphi_{xxx}^{(j)} + \frac{1}{2} \left(\sum_{i=0}^{j} \varphi^{(i)} \varphi^{(j-i)} \right)_x \right). \tag{6.5b}$$

Remember that $\varphi^{(j)}(0)=g^{(j)}(0)$, where $g^{(j)}(t)=\partial_t^j g(t)$ as before, is just the jth-order compatibility condition implied by the KdV equation (6.1a) for solutions that are sufficiently smooth at the origin $(0,0)$. Here is the main result.

THEOREM 6.2. *Let k be a positive integer, $f \in H^{3k+1}(\mathbb{R}^+)$ and $g \in H^{k+1}_{\mathrm{loc}}(\mathbb{R}^+)$. Suppose the $k+1$ compatibility conditions*

$$g^{(j)}(0) = \varphi^{(j)}(0) \quad \text{for } 0 \le j \le k,$$

hold, where $\varphi^{(j)}$ is defined above. Then there exists a unique solution u in $L^\infty_{\mathrm{loc}}(\mathbb{R}^+; H^{3k+1}(\mathbb{R}^+))$ of (6.1) *corresponding to the data f and g. In case $k>1$, u defines a classical solution, up to the boundary, of* (6.1) *in the quarter-plane $\mathbb{R}^+ \times \mathbb{R}^+$.*

The proof of this result relies on the theory for the regularized problem developed in §§3, 4, and culminating in Theorem 5.3. To make use of the last-quoted result, the following technical lemma seems essential.

LEMMA 6.3. *Let f and g be as in Theorem* 6.2. *Then there exist sequences $\{f_N\}_1^\infty \subseteq H^\infty(\mathbb{R}^+)$ and $\{g_N\}_1^\infty \subseteq C^\infty(\mathbb{R}^+)$ such that*

(i) $g_N^{(j)}(0) = \varphi_N^{(j)}(0)$ *for* $0 \le j \le k$;

(ii) $f_N \to f$ *in* $H^{3k+1}(\mathbb{R}^+)$, $g_N \to g$ *in* $H^{k+1}_{\mathrm{loc}}(\mathbb{R}^+)$.

Here $\varphi_N^{(j)}$ is as defined in (6.5) *with f_N replacing f and $g_N^{(j)} = \partial_t^j g_N$.*

Proof. Let $\{f_N\}_1^\infty \subseteq H^\infty(\mathbb{R}^+)$ and $\{h_N\}_1^\infty \subseteq C^\infty(\mathbb{R}^+)$ satisfy condition (ii) in the statement of the lemma, relative to f and g, respectively. Define

$$a_j^N = h_N^{(j)}(0) - \varphi_N^{(j)}(0) \quad \text{for } 0 \le j \le k,$$

where $h_N^{(j)} = \partial_t^j h_N$ and $\varphi_N^{(j)}$ is given as in (6.5). Then set

$$g_N(t) = h_N(t) - P_N(t),$$

where

$$P_N(t) = \sum_{j=0}^{k} a_j^N \frac{t^j}{j!}.$$

By construction, for $0 \le j \le k$,

$$g_N^{(j)}(0) = h_N^{(j)}(0) - a_j^N = \varphi_N^{(j)}(0).$$

Moreover, $g_N \in C^\infty(\mathbb{R}^+)$, for each N. It remains to verify that $g_N \to g$ in $H^{k+1}_{\text{loc}}(\mathbb{R}^+)$. This will be true if and only if $P_N \to 0$ in $H^{k+1}_{\text{loc}}(\mathbb{R}^+)$. But, for $0 \le j \le k$,

$$\lim_{N\to\infty} a_j^N = \lim_{N\to\infty} \left[h_N^{(j)}(0) - \varphi_N^{(j)}(0)\right] = 0,$$

since f and g satisfy $k+1$ compatibility conditions. Let $T>0$ be given. Then

$$\|P_N\|_{H^{k+1}(0,T)} \le \sum_{j=0}^{k} \left|a_j^N\right| \frac{1}{j!} \|t^j\|_{H^{k+1}(0,T)} \le \sum_{j=0}^{k} M_j \left|a_j^N\right|,$$

where the constants M_j depend only on j and T. Since $a_j^N \to 0$, as $N \to +\infty$, for each j, it follows that

$$\|P_N\|_{H^{k+1}(0,T)} \to 0,$$

as $N \to +\infty$. Since $T>0$ was arbitrary, the lemma is established. □

The next step in the proof of Theorem 6.2 is to establish that solutions of (6.1) exist in case f and g happen to be infinitely smooth.

PROPOSITION 6.4. *Let there be given a positive number T and a positive integer k. Let $f \in H^\infty(\mathbb{R}^+)$ and $g \in H^\infty(0,T)$ satisfy $k+1$ compatibility conditions,*

$$g^{(j)}(0) = \varphi^{(j)}(0) \quad \text{for } 0 \le j \le k.$$

Then there exists a solution u of (6.1) in $L^\infty(0,T;H^{3k+1}(\mathbb{R}^+))$ corresponding to the data f and g. Moreover, there exists a constant

$$b = b\left(\|f\|_{3k+1}, |g|_{k+1,T}\right),$$

such that

$$\|u^{(j-1)}(\cdot,t)\|_3 + \|u^{(j)}(\cdot,t)\|_1 \le b, \tag{6.6}$$

for $1 \le j \le k$, where $u^{(j)} = \partial_t^j u$. The constant b depends continuously on its arguments.

Proof. The proposition follows from Theorem 5.3. More precisely, Theorem 5.3 provides the following. There is a $\delta>0$ and a family $\{g_\varepsilon\}_{0<\varepsilon\le\delta} \subseteq H^\infty(0,T)$ such that $g_\varepsilon(0) = f(0)$, and

$$|g_\varepsilon - g|_{k+1,T} \to 0 \quad \text{as } \varepsilon \downarrow 0.$$

Let u_ε be the solution of the regularized initial- and boundary-value problem (4.1), corresponding to the data f and g_ε. Then there is a constant $b = b(\|f\|_{3k+1}, |g|_{k+1,T})$ depending continuously on its arguments, but independent of ε in $(0,\delta]$, such that

$$\begin{aligned} &\|u^{(j-1)}(\cdot,t)\|_3^2 + \varepsilon \|u_{xxxx}^{(j-1)}(\cdot,t)\|^2 + \|u^{(j)}(\cdot,t)\|_1^2 \\ &\quad + \int_0^t \left\{\left[u_{xxxx}^{(j-1)}(0,s)\right]^2 + \left[u_{xxx}^{(j-1)}(0,s)\right]^2 + \left[u_x^{(j)}(0,s)\right]^2 + \varepsilon\left[u_{xx}^{(j)}(0,s)\right]^2\right\} ds \le b, \end{aligned} \tag{6.7}$$

for $0 \le j \le k$. (In (6.7), the subscript ε has been suppressed when writing u_ε.) And, from Corollary 3.9,

$$\partial_t^i u_\varepsilon \in C(0,T;H^m(\mathbb{R}^+)),$$

for all nonnegative integers i and m. Thus

$$\left\{\partial_t^j u_\varepsilon\right\}_{0<\varepsilon\le\delta} \text{ is bounded in } L^\infty\left(0,T;H^3(\mathbb{R}^+)\right),$$

for $0 \le j < k$, and

$$\left\{\partial_t^k u_\varepsilon\right\}_{0<\varepsilon\le\delta} \text{ is bounded in } L^\infty\left(0,T;H^1(\mathbb{R}^+)\right).$$

If H is any Hilbert space, then $L^\infty(0,T;H)$ is the dual of $L^1(0,T;H)$. (Here, H is identified with its dual space.) In consequence of this fact, the unit ball in $L^\infty(0,T;H)$ is compact, for the weak-star topology induced by $L^1(0,T;H)$. Hence, by taking a sequence from $(0,\delta]$ converging to 0, and passing progressively to further subsequences, we deduce the existence of a sequence $\{\varepsilon_n\}_1^\infty$, with $\varepsilon_n \downarrow 0$ such that if

$$u_n(x,t) = u_{\varepsilon_n}(x,t), \qquad n = 1,2,3,\cdots,$$

then there are functions u and U_j in $L^\infty(0,T;H^3(\mathbb{R}^+))$, $0<j<k$, and a function U_k in $L^\infty(0,T;H^1(\mathbb{R}^+))$, such that

$$\begin{aligned} &u_n \to u && \text{weak-star in } L^\infty\big(0,T;H^3(\mathbb{R}^+)\big), \\ &\partial_t^j u_n \to U_j && \text{weak-star in } L^\infty\big(0,T;H^3(\mathbb{R}^+)\big) \quad \text{for } 0<j<k, \text{ and} \\ &\partial_t^k u_n \to U_k && \text{weak-star in } L^\infty\big(0,T;H^1(\mathbb{R}^+)\big), \end{aligned} \tag{6.8}$$

as $n \to +\infty$. Since $u_n \to u$ weak-star in $L^\infty(0,T;H^3(\mathbb{R}^+))$, certainly $u_n \to u$ in $\mathcal{D}'(0,T;H^3(\mathbb{R}^+))$. Hence $\partial_t^j u_n \to \partial_t^j u$, for all j, at least in the distributional sense. Because of (6.8), we may therefore identify U_j with $\partial_t^j u$, for $0<j\le k$.

Note also that if $\nabla u_n = (\partial_x u_n, \partial_t u_n)$, then $\{\nabla u_n\}_1^\infty$ comprises a bounded sequence in $L^\infty(0,T;H^1(\mathbb{R}^+)) \times L^\infty(0,T;H^1(\mathbb{R}^+))$. Since $H^1(\mathbb{R}^+) \subset C_b(\mathbb{R}^+)$, this means that each component of $\{\nabla u_n\}_1^\infty$ is a sequence uniformly bounded in $L^\infty(\mathbb{R}^+ \times (0,T))$. In consequence, $\{u_n\}_1^\infty$ forms an equicontinuous sequence, when restricted to any compact subset of $\overline{\mathbb{R}}^+ \times [0,T]$. Hence for any $M>0$, $\{u_n\}_1^\infty$ is precompact in $C([0,M]\times[0,T])$, by the Ascoli–Arzela lemma. So by passing to still further subsequences, and finishing off with a Cantor diagonalization, it may be presumed that

$$u_n \to u \quad \text{as } n \to +\infty, \text{ uniformly on compact subsets of } \overline{\mathbb{R}}^+ \times [0,T].$$

(More precisely, this argument leads to the conclusion that $u_n \to v$, uniformly on compact subsets of $\overline{\mathbb{R}}^+ \times [0,T]$, as $n \to +\infty$. This in turn implies that $u_n \to v$ in $\mathcal{D}'(\mathbb{R}^+ \times (0,T))$ and thus leads to the identification $v = u$.) Exactly the same argument holds good for $\partial_t^j u_n$, provided $j<k$. Thus, for $0 \le j < k$,

$$\partial_t^j u_n \to \partial_t^j u \quad \text{as } n \to +\infty, \text{ uniformly on compact subsets of } \overline{\mathbb{R}}^+ \times [0,T]. \tag{6.9}$$

By a different argument, which makes use of the fact that $H^1(0,M)$ is compactly embedded in $L_2(0,M)$ for any $M>0$ (cf. [8, Lemma 7]) it may also be presumed that

$$\partial_t^k u_n \to \partial_t^k u \quad \text{as } n \to +\infty, \text{ almost everywhere in } \overline{\mathbb{R}}^+ \times [0,T]. \tag{6.10}$$

By passing to a further subsequence, if necessary, it may be supposed as well that, as $n \to +\infty$,

$$\begin{aligned} &u_n \partial_x u_n \to w && \text{weak-star in } L^\infty\big(0,T;H^2(\mathbb{R}^+)\big), \\ &\partial_x u_n \to v && \text{weak-star in } L^\infty\big(0,T;H^2(\mathbb{R}^+)\big), \\ &\partial_x^3 u_n \to V && \text{weak-star in } L^\infty\big(0,T;L^2(\mathbb{R}^+)\big). \end{aligned}$$

Because of (6.9), $u_n \to u$ and $u_n^2 \to u^2$ in $\mathcal{D}'(\mathbb{R}^+ \times (0,T))$. Hence the identifications $w = \frac{1}{2}\partial_x u^2$, $v = \partial_x u$, $V = \partial_x^3 u$ follow. Moreover, $\partial_t \partial_x^2 u_n$ is bounded in $L^\infty(0,T;H^{-1}(\mathbb{R}^+))$, so $\varepsilon_n \partial_t \partial_x^2 u_n \to 0$ strongly in this space, as $n \to +\infty$.

The reader will now appreciate that there is in hand enough information to pass to the limit $n \to +\infty$ in the regularized equation and conclude that, at least in the

distributional sense, u satisfies the KdV equation,

$$u_t+u_x+uu_x+u_{xxx}=0,$$

in $\mathbb{R}^+\times(0,T)$. Moreover, as $u_\varepsilon(x,0)\equiv f(x)$ and $u_\varepsilon(0,t)=g_\varepsilon(t)$ for $0<\varepsilon\leq\delta$, it follows from (6.9), for example, that

$$u(x,0)=f(x) \quad \text{for } x\in\mathbb{R}^+,$$

and

$$u(0,t)=g(t) \quad \text{for } t\in[0,T].$$

Thus u does indeed provide a solution of (6.1) on $\overline{\mathbb{R}}^+\times[0,T]$. Moreover, by the lower-semicontinuity of the norm, relative to weak-star convergence, (6.7) implies that

$$\|u^{(j)}(\cdot,t)\|_3\leq b,$$

for $0\leq j<k$, and

$$\|u^{(k)}(\cdot,t)\|_1\leq b,$$

where $b=b(\|f\|_{3k+1},|g|_{k+1,T})$ is the constant obtained earlier from Theorem 5.3.

Notice that, if $k=1$, then $u_t\in L^\infty(0,T;H^1(\mathbb{R}^+))$ and u_x, $uu_x\in L^\infty(0,T;H^2(\mathbb{R}^+))$. Hence, from the differential equation, $u_{xxx}\in L^\infty(0,T;H^1(\mathbb{R}^+))$, whence $u\in L^\infty(0,T;H^4(\mathbb{R}^+))$. If $k>1$, this type of simple argument may be continued inductively. The outcome is that

$$\partial_t^j u\in L^\infty\big(0,T;H^{3(k-j)+1}(\mathbb{R}^+)\big), \tag{6.11}$$

for $0\leq j\leq k$.

Finally, (6.11) and standard interpolation results ([19, Chap. 1, Thm. 3.1]) yield the following additional smoothness results:

$$\partial_t^j u\in C\big(0,T;H^{3(k-j)-1/2}(\mathbb{R}^+)\big), \tag{6.12}$$

for $0\leq j<k$.

In particular, if $k>1$, certainly $u\in C(0,T;H^4(\mathbb{R}^+))$. Therefore, u_t, u_x, uu_x, and u_{xxx} all lie in $C(0,T;H^1(\mathbb{R}^+))$. As this latter space is embedded in $C_b(\overline{\mathbb{R}}^+\times[0,T])$, it follows that, after possible modification on a set of measure zero, all the derivatives in the differential equation are continuous, and bounded, functions. Consequently, if $k>1$, u is a classical solution of the quarter-plane problem for KdV.

The proof of the proposition is now completed. □

Remark. Because the solution u obtained in Proposition 6.4 lies within the realm of the uniqueness theorem 6.1, the entire family $\{u_\varepsilon\}_{0<\varepsilon\leq\delta}$ is inferred to converge to u, in the various senses appearing in the proof. This is because we actually prove that any sequence $\{\varepsilon_n\}_1^\infty$ in $(0,\delta]$, with $\varepsilon_n\to 0$, as $n\to+\infty$, has a subsequence such that the corresponding functions $\{u_n\}$ converge to a solution of (6.1), which by uniqueness must be u.

The last proposition gives very nearly the result stated in Theorem 6.2. The only essential difference is that f and g are assumed to be infinitely differentiable. Using Lemma 6.3, this added assumption is shown to be unnecessary.

Proof of Theorem 6.2. Suppose now that $f\in H^{3k+1}(\mathbb{R}^+)$ and $g\in H^{k+1}_{\text{loc}}(\mathbb{R}^+)$ are fixed, and that f and g satisfy the first $k+1$ compatibility conditions, as in the

statement of the theorem. Fix $T>0$. By Lemma 6.3, there exist sequences $\{f_N\}_1^\infty \subseteq H^\infty(\mathbb{R}^+)$ and $\{g_N\}_1^\infty \subseteq C^\infty(\mathbb{R}^+)$ such that

$$f_N \to f \quad \text{in } H^{3k+1}(\mathbb{R}^+), \tag{6.13}$$
$$g_N \to g \quad \text{in } H^{k+1}(0,T),$$

as $N \to +\infty$. And, for each $N>0$, f_N and g_N satisfy the same $k+1$ compatibility conditions satisfied by f and g. The last proposition thus applies, and it is concluded that there is a solution u_N of (6.1), on $\overline{\mathbb{R}}^+ \times [0,T]$, corresponding to the data f_N and g_N. Moreover, $\partial_t^j u_N \in L^\infty(0,T; H^{3(k-j)+1}(\mathbb{R}^+))$, for $0 \le j \le k$, and if

$$b_N = b\big(\|f_N\|_{3k+1}, |g_N|_{k+1,T}\big),$$

then for $0 \le j < k$,

$$\|\partial_t^j u_N\|_{L^\infty(0,T;H^3(\mathbb{R}^+))} \le b_N, \qquad \|\partial_t^k u_N\|_{L^\infty(0,T;H^1(\mathbb{R}^+))} \le b_N.$$

Because of (6.13) and the fact that b is bounded as its arguments vary over a bounded set, there is a constant B, independent of N, such that

$$\|\partial_t^j u_N\|_{L^\infty(0,T;H^3(\mathbb{R}^+))} \le B, \tag{6.14a}$$

for $0 \le j < k$, and

$$\|\partial_t^k u_N\|_{L^\infty(0,T;H^1(\mathbb{R}^+))} \le B. \tag{6.14b}$$

In consequence of the bounds expressed in (6.14), the arguments of Proposition 6.4 may be repeated without essential change (the extra smoothness available during the proof of the proposition was not used, nor was the regularizing term $-\varepsilon u_{xxt}$). It is concluded therefore that $\{u_N\}_1^\infty$ converges to a function u_T, say, in the various ways already detailed in the proof of Proposition 6.4. As before, u_T provides a solution of (6.1) corresponding to the data f and g, on $\overline{\mathbb{R}}^+ \times [0,T]$.

The above argument applies for any fixed $T>0$. Define a function U on $\mathbb{R}^+ \times \mathbb{R}^+$ by,

$$U(x,t) = u_T(x,t),$$

provided that $t<T$. This is well defined because of the uniqueness result. It is clear that U provides the solution whose existence was contemplated in the statement of Theorem 6.2. The fact that U is a classical solution of the problem (6.1), if $k>1$, follows exactly as in the proof of Proposition 6.4. The theorem is thus established. □

It is perhaps worth comment that Theorem 6.2 also holds if $k=0$. This result subsists on the ε-independent $H^1(\mathbb{R}^+)$-bound established in Corollary 3.6. The proof of existence of these weaker solutions, while a little more delicate than the proof of Theorem 6.2, fits more or less directly into the framework exposed in the proof of Proposition 6.3. (The extra ingredients may be found, for example, in [8, App. A].) For this reason, we content ourselves with a statement of this further consequence.

THEOREM 6.5. *Let $f \in H^1(\mathbb{R}^+)$ and $g \in H^1_{\text{loc}}(\mathbb{R}^+)$, and suppose $f(0)=g(0)$. Then there exists a solution u in $L^\infty_{\text{loc}}(\mathbb{R}^+; H^1(\mathbb{R}^+))$ of problem (6.1) corresponding to the data f and g.*

Remarks. By a solution we mean as usual a solution in the sense of distributions. In this case the uniqueness result does not apply.

Note that, for any $T>0$, $u_t \in L^\infty(0,T; H^{-2}(\mathbb{R}^+))$, from the equation. Hence $u \in C(0,T; H^{-1/2}(\mathbb{R}^+))$ (cf. again [19, Chap. 1]), so the initial-value is taken on in a weak,

but meaningful way. Note as well that $L^\infty(0,T;H^1(\mathbb{R}^+)) \subseteq L^\infty(0,T;C_b(\mathbb{R}^+))$. Hence for almost every t in $[0,T]$, $u(x,t)$ is continuous in x at $x=0$. Thus the boundary-values are also obtained in a meaningful way.

7. Conclusion. The quarter-plane problem (1.3) is argued to be a natural configuration in which to use the KdV equation for the prediction of wave propagation in a uniform channel. The general idea behind the use of this form of initial- and boundary-value problem for testing the appurtenance of the KdV equation may be appreciated by reference to Fig. 1. With the liquid initially at rest ($f \equiv 0$), a wavemaker located at one end of the channel is activated. The passage of the waves down the channel is recorded by probes, the recording nearest the wavemaker being construed as the boundary data $g(t)$. Note that if the waves are in the regime to which, formally, KdV applies, then they are expected to be smooth, and so g will lie in $\mathcal{D}(0,T)$, for some $T>0$. In consequence, the data so determined will satisfy the compatibility conditions, expressed for example below (6.5), to all orders. Hence the theory developed herein is applicable.

Our theory demonstrates that problem (1.3) has unique smooth solutions corresponding to such smooth and compatible data. This is a step in the direction of a satisfactory mathematical analysis of the situation envisaged in Fig. 1. Another important step, which has not been treated here, is a result of continuous dependence of the solutions on variations of the data. Also, in considering comparisons of the model's predictions with laboratory-scale experiments, some compensation for dissipative effects must be included (cf. [10]). Less important, but still of some mathematical interest, is a possible improvement of the regularity theory to bring this aspect into line with the theory for the pure initial-value problem (cf. [8] or [16]). We have shown that if $f \in H^{3k+1}(\mathbb{R}^+)$ and $g \in H^{k+1}_{\text{loc}}(\mathbb{R}^+)$ satisfy the appropriate compatibility conditions at $(x,t)=(0,0)$, then the quarter-plane problem has a solution in $L^\infty_{\text{loc}}(\mathbb{R}^+;H^{3k+1}(\mathbb{R}^+))$. Whereas, we confidently expect the solutions to lie in $C(\mathbb{R}^+;H^{3k+1}(\mathbb{R}^+))$. In fact, this latter point seems to be related to a sharp version of continuous dependence of solutions on the data.

It deserves emphasis that a satisfactory numerical scheme for the configuration in view here is essential to effect any quantitative comparisons of laboratory data with predictions of the model. Especial care must be exercised here. First, control of the high-frequency end of the Fourier spectrum must be assured. Otherwise an untenable error may be created near $x=0$, due to the large negative phase and group velocity associated to such components (cf. [4, §2]). Secondly, the integration will in fact take place on a bounded spatial domain, forcing the imposition of additional boundary conditions. This in turn will lead to consideration of an initial- and two-point-boundary-value problem for the KdV equation, and to consideration of the relation of such a problem to the situation studied here. The difficulties seem numerous enough to warrant insisting on a scheme having rigorously derived error bounds. Thus far, such schemes seem to be available only for the periodic initial-value problem (cf. [1], [2], [29] and [30]).

Finally, it is worth remarking that the methods embodied in this paper might yield a comparison theorem between the quarter-plane problem (1.3) for KdV and the analogous quarter-plane problem for (1.4) studied in [5], and used in the comparisons with experimental data reported in [10]. Such a program of comparison of model equations has been carried out for the associated pure initial-value problems in [11], using the general line pursued herein. Thus there is some cause for hope that a similar result is obtained in the present context.

Acknowledgments. The first author wishes to record stimulating and useful conversations and correspondence with M. Heard, W. G. Pritchard, L. R. Scott and R. Smith concerning the problem studied herein. Both authors gratefully acknowledge support from the National Science Foundation during part of this collaboration. Both authors were also supported by the U. S. Army Research Office as visiting members of the Mathematics Research Center, University of Wisconsin-Madison, during part of this collaboration.

REFERENCES

[1] D. N. ARNOLD AND R. WINTHER, *A superconvergent finite element method for the Korteweg–de Vries equation*, Math. Comp., 38, (1982), pp. 23–36.

[2] G. A. BAKER, V. A. DOUGALIS AND O. A. KARAKASHIAN, *Convergence of Galerkin approximations for the Korteweg–de Vries equation*, Math. Comp., to appear.

[3] T. B. BENJAMIN, *Lectures on nonlinear wave motion*, in Lectures in Applied Mathematics, vol. 15, A. Newell, ed., American Mathematical Society, Providence, RI, 1974.

[4] T. B. BENJAMIN, J. L. BONA AND J. J. MAHONY, *Model equations for long waves in nonlinear dispersive systems*, Phil. Trans. Roy. Soc. London A, 272, (1972), pp. 47–78.

[5] J. L. BONA AND P. J. BRYANT, *A mathematical model for long waves generated by wavemakers in non-linear dispersive systems*, Proc. Camb. Phil. Soc., 73, (1973), pp. 391–405.

[6] J. L. BONA AND M. HEARD, *An application of the general theory for quasi-linear evolution equations to an initial- and boundary-value problem for the Korteweg–de Vries equation*, in preparation.

[7] J. L. BONA AND M. E. SCHONBEK, *Long-wave models with initial data corresponding to bore propagation*, in preparation.

[8] J. L. BONA AND R. SMITH, *The initial-value problem for the Korteweg–de Vries equation*, Phil. Trans. Roy. Soc. London A, 278, (1975), pp. 555–601.

[9] _____, *A model for the two-way propagation of water waves in a channel*, Math. Proc. Camb. Phil. Soc., 79, (1976), pp. 167–182.

[10] J. L. BONA, W. G. PRITCHARD AND L. R. SCOTT, *An evaluation of a model equation for water waves*, Phil. Trans. Roy Soc. London A, 302, (1981), pp. 457–510.

[11] _____, *A comparison of solutions of model equations for long waves*, in Lectures in Applied Mathematics, vol. 20, N. Lebovitz, ed., American Mathematical Society, Providence, RI, 1983.

[12] J. L. HAMMACK, *A note on tsunamis: their generation and propagation in an ocean of uniform depth*, J. Fluid Mech., 60, (1973), pp. 769–799.

[13] J. L. HAMMACK AND H. SEGUR, *The Korteweg–de Vries equation and water waves. 2. Comparison with experiments*, J. Fluid Mech., 65, (1974), pp. 289–314.

[14] A. JEFFREY AND T. KAKUTANI, *Weak nonlinear dispersive waves, a discussion centred around the Korteweg–de Vries equation*, SIAM Rev., 14, (1972), pp. 582–643.

[15] T. KAKUTANI AND K. MATSUUCHI, *Effect of viscosity on long gravity waves*, J. Phys. Soc. Japan 39, (1975), pp. 237–246.

[16] T. KATO, *Quasi-linear equations of evolution with applications to partial differential equations*, Lecture Notes in Mathematics, 448, Springer, New York, 1975, pp. 25–70.

[17] P. D. LAX, *Almost periodic solutions of the KdV equation*, SIAM Rev., 18, (1976), pp. 351–375.

[18] J. L. LIONS, *Quelques méthodes de resolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[19] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York, 1972.

[20] A. MENIKOFF, *Unbounded solutions of the Korteweg–de Vries equation*, Comm. Pure Appl. Math., 25, (1972), pp. 407–432.

[21] R. M. MIURA, *The Korteweg–de Vries equation: a model for nonlinear dispersive waves*, in Nonlinear Waves, S. Leibovich and R. Seebass, eds., Cornell Univ. Press, Ithaca, NY, 1974.

[22] _____, *The Korteweg–de Vries equation: a survey of results*, SIAM Rev., 18, (1976), pp. 412–459.

[23] D. H. PEREGRINE, *Calculations of the development of an undular bore*, J. Fluid Mech., 25, (1966), pp. 321–330.

[24] A. C. SCOTT, F. Y. F. CHU AND D. W. MCLAUGHLIN, *The soliton: a new concept in applied science*, Proc. IEEE, 61, (1973), pp. 1443–1483.

[25] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton Univ. Press, Princeton, NJ, 1970.

[26] R. TEMAM, *Sur une problème non linéaire*, J. Math. Pures Appl., 48, (1969), pp. 159–172.
[27] B. A. TON, *Initial boundary-value problems for the Korteweg–de Vries equation*, J. Differential Equations, 25, (1977), pp. 288–309.
[28] F. TREVES, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.
[29] L. WAHLBIN, *Method for the numerical solution of first order hyperbolic equations*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974.
[30] R. WINTHER, *A conservative finite element method for the Korteweg–de Vries equation*, Math. Comp., 34, (1980), pp. 23–43.
[31] N. J. ZABUSKY AND C. J. GALVIN, *Shallow-water waves, the Korteweg–de Vries equation and solitons*, J. Fluid Mech., 47, (1971), pp. 811–824.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0004 $01.25/0

A FREE BOUNDARY PROBLEM ARISING FROM A BISTABLE REACTION-DIFFUSION EQUATION*

DAVID TERMAN†

Abstract. The pure initial value problem for the bistable reaction-diffusion equation

$$v_t = v_{xx} + f(v)$$

is considered. Here $f(v)$ is given by $f(v) = -v + H(v-a)$ where H is the Heaviside step function, and $a \in (0, \frac{1}{2})$. It is demonstrated that this equation exhibits a threshold phenomenon. This is done by considering the curve $s(t)$ defined by $s(t) = \sup\{x : v(x,t) = a\}$. It is shown that if $v(x,0) < a$ for all x, then $\lim_{t\to\infty} \|v(\cdot,t)\|_\infty = 0$. Moreover, there exists a positive constant c^* such that if the initial datum is sufficiently smooth and satisfies $v(x,0) > a$ on a sufficiently long interval, then $s(t)$ is defined in $\mathbb{R}^+$, and $\lim_{t\to\infty}(s(t) - c^*t)$ exists. Regularity and uniqueness properties of $s(t)$ are also presented.

AMS-MOS subject classification (1980). Primary 35K55

Key words. reaction-diffusion equations, threshold phenomena, free boundary problem

1. Introduction. In this paper we consider the pure initial value problem for the equation

$$v_t = v_{xx} + f(v), \tag{1.1}$$

the initial datum being $v(x,0) = \varphi(x)$. We assume that $f(v) = -v + H(v-a)$ where H is the Heaviside step function and $a \in (0, \frac{1}{2})$. This equation, but with smooth f, has many applications and has been studied by a number of authors (see [1], [3], [4], [7], [8], [9]). Equation (1.1) is also a special case of the FitzHugh–Nagumo equations:

$$\begin{aligned} v_t &= v_{xx} + f(v) - w, \\ w_t &= \varepsilon(v - \gamma w), \qquad \varepsilon \geq 0, \quad \gamma \geq 0, \end{aligned} \tag{1.2}$$

which were introduced as a model for the conduction of electrical impulses in the nerve axon. Note that (1.1) can be obtained from (1.2) by setting $\varepsilon = 0$ and $w \equiv 0$ in $\mathbb{R} \times \mathbb{R}^+$. In their original model, FitzHugh [5] and Nagumo, et al. [11] chose $f(v) = v(1-v)(v-a)$. McKean [10] suggested the further simplification $f(v) = -v + H(v-a)$.

Our primary interest is to study the asymptotic behavior of solutions of equation (1.1). One expects equation (1.1) to exhibit a threshold phenomenon. That is, if the initial datum $\varphi(x)$ is sufficiently small then one expects the solutions of (1.1) to decay exponentially fast to zero as $t \to \infty$. This corresponds, for example, to the biological fact that a minimum stimulus is needed to trigger a nerve impulse. In this case we say that $\varphi(x)$ is subthreshold. One expects, however, that if $\varphi(x)$ is sufficiently large, or superthreshold, then some sort of signal will propagate. Threshold results for equation (1.1) with smooth "cubic-like" f have been given by Aronson and Weinberger [1]. Fife and McLeod [3] showed that if the initial datum is superthreshold, then the solution of (1.1), with smooth f, will converge to a traveling wave solution.

*Received by the editors November 20, 1981, and in revised form August 25, 1982. This material is based upon work supported by the National Science Foundation under grant MCS80-17158, and sponsored by the U. S. Army under contract DAAG29-80-C-0041.

†Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

Throughout this paper we assume that the initial datum, $\varphi(x)$, satisfies the following conditions:

(1.3) (a) $\varphi(x)\in C^1(\mathbb{R})$,
(b) $\varphi(x)\in[0,1]$ in $\mathbb{R}$,
(c) $\varphi(x)=\varphi(-x)$ in $\mathbb{R}$,
(d) $\varphi'(x)<0$ in $\mathbb{R}^+$,
(e) $\varphi(x_0)=a$ for some $x_0>0$,
(f) $\varphi''(x)$ is a bounded, continuous function except possibly at $|x|=x_0$.

This last condition is needed in order to obtain sufficient a priori bounds on the derivatives of the solution of (1.1).

Note that in some sense x_0 determines the size of the initial datum. We expect, therefore, a signal to propagate if x_0 is sufficiently large. In order to be more precise we consider the curve $s(t)$ given by

$$s(t)=\sup\{x : v(x,t)=a\}. \tag{1.4}$$

We say that the initial datum is superthreshold if $s(t)$ is defined in $\mathbb{R}^+$ and $\lim_{t\to\infty} s(t)=+\infty$. In this paper we show that if x_0 is sufficiently large then $\varphi(x)$ is indeed superthreshold. We also analyze the asymptotic behavior of the curve $s(t)$. We prove that there exists a constant c^* such that if $\varphi(x)$ is superthreshold, then $\lim_{t\to\infty}(s(t)-c^*t)$ exists. That is the solution eventually propagates with constant velocity.

Note that because $f(v)$ is discontinuous we cannot expect the solution of equation (1.1) to be very smooth. By a classical solution of equation (1.1) we mean the following:

DEFINITION. Let $S_T=\mathbb{R}\times(0,T)$ and $G_T=\{(x,t)\in S_T,\ v(x,t)\neq a\}$. Then $v(x,t)$ is said to be a classical solution of the Cauchy problem (1.1) in S_T if

(a) v, along with v_x, are bounded continuous functions in S_T,

(b) in G_T, v_{xx} and v_t are continuous functions which satisfy the equation

$$v_t=v_{xx}+f(v),$$

(c) $\lim_{t\downarrow 0} v(x,t)=\varphi(x)$ for each $x\in\mathbb{R}$.

We can now state our primary result.

THEOREM 1.1. *Choose $a\in(0,\frac{1}{2})$. Then there exist positive constants θ and c^* such that if $\varphi(x)$ satisfies the conditions* (1.3) *with $x_0>\theta$, then* (1.1) *possesses a classical solution in $\mathbb{R}\times\mathbb{R}^+$, and $\varphi(x)$ is superthreshold. Furthermore,*

(a) $s(t)\in C^1(\mathbb{R}^+)$,

(b) $s'(t)$ *is a locally Lipschitz continuous function,*

(c) $\lim_{t\to\infty}(s(t)-c^*t)$ *exists.*

Actually, assumption (1.3c) is not needed in the proof that $\lim_{t\to\infty}(s(t)-c^*t)$ exists. All we need assume is that $s(t)\in C^1(\mathbb{R}^+)$ and $\lim_{t\to\infty} s(t)=\infty$.

Theorem 1.1 plays an essential role in a recent paper [15] in which the author proves a threshold result for the full FitzHugh–Nagumo system. That result is proved by showing that the variable $v(x,t)$ in (1.2) lies above some comparison function, $u(x,t)$, which is essentially the solution of a scalar equation of the form (1.1). If we define $\sigma(t)$ by $\sigma(t)=\sup\{x : u(x,t)=a\}$ then the properties needed about $\sigma(t)$ in [15] follow from Theorem 1.1. In particular, in order to apply the basic comparison theorems it is crucial that $\sigma(t)$ is sufficiently smooth. In order to prove threshold results it is needed that $\lim_{t\to\infty}\sigma(t)=\infty$ if the initial datum is sufficiently large.

Note that for the model we are considering it is trivial to give sufficient conditions for the initial datum to be subthreshold. In particular, if $\varphi(x)<a$ for each $x\in\mathbb{R}$ then, from the maximum principle (see [12, p. 159]), $v(x,t)<a$ in $\mathbb{R}\times\mathbb{R}^+$. Hence v satisfies the equation

$$v_t=v_{xx}-v \quad \text{in } \mathbb{R}\times\mathbb{R}^+.$$

From this it follows that $\|v(\cdot,t)\|_\infty\to 0$ as $t\to\infty$, and the initial datum is subthreshold.

To prove Theorem 1.1 we first demonstrate that if the initial datum $\varphi(x)$ satisfies the conditions (1.3) then there must exist some positive T such that in the interval $[0,T]$, $s(t)$ satisfies the integral equation

$$a-\int_{-\infty}^{\infty}K(s(t)-\xi,t)\varphi(\xi)\,d\xi=\int_0^t d\tau\int_{-s(\tau)}^{s(\tau)}K(s(t)-\xi,t-\tau)\,d\xi \tag{1.5}$$

where $K(x,t)=(e^{-t}/2\pi^{1/2}t^{1/2})e^{-x^2/4t}$ is the fundamental solution of the linear differential equation $\psi_t=\psi_{xx}-\psi$. Here we give a formal explanation of why this is true. We then show how to construct a solution of the initial value problem (1.1) given a smooth solution of (1.5).

From assumptions (1.3c, d) we expect that $v_x(x,t)<0$ in $\mathbb{R}^+\times\mathbb{R}^+$. In this case $s(t)$ will be a well defined, continuous function for some time, say $t\in[0,T]$. It also follows that $v>a$ for $|x|<s(t)$ and $v<a$ for $|x|>s(t)$. Let $\chi_{\mathcal{G}}$ be the indicator function of the set $\mathcal{G}=\{(x,t):v(x,t)>a;\ 0\le t\le T\}$. Then, for $|x|\ne s(t)$, $v(x,t)$ satisfies the inhomogeneous equation

$$v_t=v_{xx}-v+\chi_{\mathcal{G}} \tag{1.6}$$

with initial datum $v(x,0)=\varphi(x)$. Formally the solution of (1.6) can be written as

$$v(x,t)=\int_{-\infty}^{\infty}K(x-\xi,t)\varphi(\xi)\,d\xi+\int_0^t d\tau\int_{-s(\tau)}^{s(\tau)}K(x-\xi,t-\tau)\,d\xi. \tag{1.7}$$

Setting $x=s(t)$ in (1.7) we obtain (1.5).

LEMMA 1.2. *Suppose that $s(t)$ is a continuously differentiable function which satisfies the integral equation* (1.5) *in* $[0,T]$. *Then the function $v(x,t)$ given by* (1.7) *is a classical solution of the initial value problem* (1.1) *in* $\mathbb{R}\times[0,T]$.

Proof. Setting $x=s(t)$ in (1.7) and subtracting the resulting equation from (1.5) we find that $v(s(t),t)=a$ in $[0,T]$. Differentiating both sides of (1.7) we see that for $x\ne s(t)$, $v(x,t)$ satisfies the differential equation $v_t=v_{xx}+f(v)$ in $\mathbb{R}\times(0,T]$. It also follows from (1.7) that $\lim_{t\downarrow 0}v(x,t)=\varphi(x)$ for $x\in\mathbb{R}$. We now show that $v(x,t)$ is differentiable whenever $x=s(t)$.

First assume that $|\xi|<s(\tau)$. Then $v(\xi,\tau)$ satisfies the differential equation

$$v_\tau-v_{\xi\xi}+v=1.$$

Multiplying both sides of this equation by $K(x-\xi,t-\tau)$ and using the fact that $K_\tau+K_{\xi\xi}-K=0$ we find that

$$(Kv)_\tau-(Kv_\xi)_\xi+(K_\xi v)_\xi=K.$$

Assuming that $|x|<s(t)$ we integrate this last equation for $-s(\tau)<\xi<s(\tau)$, $\varepsilon<\tau<t-\varepsilon$, and let $\varepsilon\to 0$ to obtain:

$$\begin{aligned}(1.8a)\quad v(x,t)&-\int_{-x_0}^{x_0}K(x-\xi,t)\varphi(\xi)\,d\xi-\int_0^t K(x-s(\tau),t-\tau)as'(\tau)\,d\tau\\&-\int_0^t K(x+s(\tau),t-\tau)as'(\tau)\,d\tau-\int_0^t K(x-s(\tau),t-\tau)v_\xi\big(s(\tau)^-,\tau\big)\,d\tau\\&+\int_0^t K(x+s(\tau),t-\tau)v_\xi\big(-s(\tau)^+,\tau\big)\,d\tau+\int_0^t aK_\xi(x-s(\tau),t-\tau)\,d\tau\\&-\int_0^t aK_\xi(x+s(\tau),t-\tau)\,d\tau\\&=\int_0^t d\tau\int_{-s(\tau)}^{s(\tau)}K(x-\xi,t-\tau)\,d\xi.\end{aligned}$$

Next assume that $\xi>s(\tau)$. Then $v(\xi,\tau)$ satisfies the differential equation: $v\tau-v_{\xi\xi}+v=0$. Multiplying both sides of this equation by $K(x-\xi,t-\tau)$ we find that

$$(Kv)_\tau-(Kv_\xi)_\xi+(K_\xi v)_\xi=0.$$

We integrate this equation for $s(\tau)<\xi<\infty$, $\varepsilon<\tau<t-\varepsilon$ and let $\varepsilon\to 0$ to obtain

$$\begin{aligned}(1.8b)\qquad&-\int_{-x_0}^{\infty}K(x-\xi,t)\varphi(\xi)\,d\xi+\int_0^t K(x-s(\tau),t-\tau)as'(\tau)\,d\tau\\&+\int_0^t K(x-s(\tau),t-\tau)v_\xi\big(s(\tau)^+,\tau\big)\,d\tau-\int_0^t aK_\xi(x-s(\tau),t-\tau)\,d\tau=0.\end{aligned}$$

Similarly, for $\xi<s(\tau)$ we obtain

$$\begin{aligned}(1.8c)\quad&-\int_{-\infty}^{-x_0}K(x-\xi,t-\tau)\,d\xi+\int_0^t K(x+s(\tau),t-\tau)as'(\tau)\,d\tau\\&-\int_0^t K(x+s(\tau),t-\tau)v_\xi\big(-s(\tau)^-,\tau\big)\,d\tau+\int_0^t aK_\xi(x+s(\tau),t-\tau)\,d\tau=0.\end{aligned}$$

Adding (1.8a), (1.8b), and (1.8c), and using (1.7) we find that for $t\in(0,T)$

$$\begin{aligned}(1.9)\qquad&\int_0^t\Big[K(x-s(\tau),t-\tau)\big[v_\xi\big(s(\tau)^+,\tau\big)-v_\xi\big(s(\tau)^-,\tau\big)\big]\\&+K(x+s(\tau),t-\tau)\big[v_\xi\big(-s(\tau)^+,\tau\big)-v_\xi\big(-s(\tau)^-,\tau\big)\big]\Big]\,d\tau=0.\end{aligned}$$

However, because of assumption (1.3c) it follows from (1.7) that $v(x,t)=v(-x,t)$ in $\mathbb{R}\times(0,T)$. Therefore, (1.9) can be rewritten as

$$\int_0^t\big[K(x-s(\tau),t-\tau)-K(x+s(\tau),t-\tau)\big]\big[v_\xi\big(s(\tau)^+,\tau\big)-v_\xi\big(s(\tau)^-,\tau\big)\big]\,d\tau=0.$$

From this it follows that $v_x(s(t)^-,t)=v_x(s(t)^+,t)$ for each $t\in(0,T)$]. □

In §2 we present some notation and prove a few preliminary results which are needed throughout the rest of the paper. In §3 we show that for some time T there exists a solution of the integral equation (1.5) in $[0,T]$. We also demonstrate that $s(t)\in C^1(0,T)$ and $s'(t)$ is a locally Lipschitz continuous function. From the proof of these results it will be clear that we may choose $T=+\infty$ if x_0 is sufficiently large. In §4 we prove that the solution of (1.5) is unique among Lipschitz continuous functions. Finally, in §5 we show that there exists constants θ and c^*, which depends only on the parameter a, such that if $x_0>\theta$, then $\lim_{t\to\infty}(s(t)-c^*t)$ exists.

2. The operators Φ and Θ. We first introduce the following notation. Assume that $\psi(x,t)$ is the solution of the linear differential equation

$$\psi_t=\psi_{xx}-\psi \tag{2.1}$$

in $\mathbb{R}\times\mathbb{R}^+$ with initial conditions

$$\psi(x,0)=\varphi(x).$$

Note that $\psi(x,t)=\int_{-\infty}^{\infty}K(x-\xi,t)\varphi(\xi)\,d\xi$ where $K(x,t)$ is the fundamental solution of (2.1).

Now suppose that $\alpha(t)$ is a positive, continuous function defined for $t\in[0,T]$. For values of t_0 and t which satisfy $0\le t_0<t\le T$ we define the operators:

$$\Phi(\alpha)(t)=\int_0^t d\tau\int_{-\alpha(\tau)}^{\alpha(\tau)}K(\alpha(t)-\xi,t-\tau)\,d\xi,$$
$$\Phi_{t_0}(\alpha)(t)=\Phi(\alpha)(t)-\Phi(\alpha)(t_0),$$
$$\Theta(\alpha)(t)=a-\psi(\alpha(t),t),$$
$$\Theta_{t_0}(\alpha)(t)=\theta(\alpha)(t)-\theta(\alpha)(t_0)=\psi(\alpha(t_0),t_0)-\psi(\alpha(t_1),t_1).$$

Note that $s(t)$ is a solution of the integral equation (1.5) in $[0,T]$ if and only if

$$\Theta_{t_0}(s)(t)=\Phi_{t_0}(s)(t)$$

for all values of t_0 and t such that $0\le t_0<t\le T$.

DEFINITION. Suppose that $\alpha(t)$ is a positive uniformly Lipschitz continuous function defined in $[0,T]$. We define $\alpha(t)$ to be a *lower solution* in $[0,T]$ if $\Phi(\alpha)(t)\ge\Theta(\alpha)(t)$ in $[0,T]$. If $\Phi(\alpha)(t)\le\Theta(\alpha)(t)$ in $[0,T]$ then $\alpha(t)$ is said to be an *upper solution* in $[0,T]$.

In Theorem 4.1 it is shown that if $\alpha(t)$ and $\beta(t)$ are respectively lower and upper solutions in $[0,T]$ then $\alpha(t)\le\beta(t)$ in $[0,T]$. This will imply that the solution of (1.5) is unique among uniformly Lipschitz functions. We prove threshold results by showing that if x_0 is sufficiently large then some vertical line $l_1(t)=\bar{x}$ is a lower solution in $\mathbb{R}^+$. Hence $s(t)\ge\bar{x}$ in $\mathbb{R}^+$. Using this preliminary result we then show that $\lim_{t\to\infty}s(t)=\infty$.

In the rest of this section we prove those properties of the operators Θ and Φ which are needed for the proof of Theorem 1.1. We assume throughout this section that $\alpha(t)$ and $\beta(t)$ are positive continuous functions defined on an interval $[0,T]$.

LEMMA 2.1. *Assume that for* $t_0<t_1$, $\alpha(t_0)\le\beta(t_0)$, *and* $\alpha(t_1)>\beta(t_1)$. *Then* $\Theta_{t_0}(\alpha)(t_1)>\Theta_{t_0}(\beta)(t_1)$.

Proof. Recall that $\theta_{t_0}(\alpha)(t_1)=\psi(\alpha(t_0),t_0)-\psi(\alpha(t_1),t_1)$ where $\psi(x,t)$ is the solution of the linear differential equation

$$\psi_t=\psi_{xx}-\psi$$

with initial datum $\psi(x,0)=\varphi(x)$. From assumption (1.3d) and the comparison theorem (see [12, p. 159]) applied to $\psi_x(x,t)$ it follows that $\psi_x(x,t)<0$ in $\mathbb{R}\times\mathbb{R}^+$. Therefore, $\psi(\alpha(t_0),t_0)>\psi(\beta(t_0),t_0)$ and $\psi(\alpha(t_1),t_1)<\psi(\beta(t_1),t_1)$. From this the proof of the lemma follows immediately. □

LEMMA 2.2. *Assume that* $\alpha(t)\ge\beta(t)$ *in* $[0,t_0]$, $\alpha(t)>\beta(t)$ *for some* $t\in(0,t_0)$, *and* $\alpha(t_0)=\beta(t_0)$. *Then* $\Phi(\alpha)(t_0)>\Phi(\beta)(t_0)$.

Proof. This is an immediate consequence of the definition of Φ. □

LEMMA 2.3. *Assume that* $\alpha(t)\in C^1(0,T)$. *Then* $\Phi(\alpha)(t)\in C^1(0,T)$ *and*

$$\Phi(\alpha)'(t)=\int_{-x_0}^{x_0}K(\alpha(t)-\xi,t)\,d\xi+\int_0^t K(\alpha(t)+\alpha(\tau),t-\tau)[\alpha'(\tau)+\alpha'(t)]\,d\tau + \int_0^t K(\alpha(t)-\alpha(\tau),t-\tau)[\alpha'(\tau)-\alpha'(t)]\,d\tau. \tag{2.2}$$

Proof. Note that

$$\begin{aligned}\Phi(\alpha)'(t)&=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[\Phi(\alpha)(t+\varepsilon)-\Phi(\alpha)(t)]\\ &=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\int_0^{t+\varepsilon}d\tau\int_{-\alpha(\tau)}^{\alpha(\tau)}K(\alpha(t+\varepsilon)-\xi,t+\varepsilon-\tau)\,d\xi - \int_0^t d\tau\int_{-\alpha(\tau)}^{\alpha(\tau)}K(\alpha(t)-\xi,t-\tau)\,d\xi\right]\\ &=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\int_{-\varepsilon}^{t}d\tau\int_{-\alpha(\tau)+\alpha(t)-\alpha(t+\varepsilon)}^{\alpha(\tau+\varepsilon)+\alpha(t)-\alpha(t+\varepsilon)}K(\alpha(t)-\xi,t-\tau)\,d\xi - \int_0^t d\tau\int_{-\alpha(\tau)}^{\alpha(\tau)}K(\alpha(t)-\xi,t-\tau)\,d\xi\right]\\ &=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\int_{-\varepsilon}^{0}d\tau\int_{-\alpha(\tau)+\alpha(t)-\alpha(t+\varepsilon)}^{\alpha(\tau+\varepsilon)+\alpha(t)-\alpha(t+\varepsilon)}K(\alpha(t)-\xi,t-\tau)\,d\xi + \int_0^t d\tau\int_{-\alpha(\tau)+\alpha(t)-\alpha(t+\varepsilon)}^{-\alpha(\tau)}K(\alpha(t)-\xi,t-\tau)\,d\xi + \int_0^t d\tau\int_{\alpha(\tau)}^{\alpha(\tau+\varepsilon)+\alpha(t)-\alpha(t+\varepsilon)}K(\alpha(t)-\xi,t-\tau)\,d\xi\right].\end{aligned}$$

Passing to the limit we obtain (2.2). □

We conclude this section by finding sufficient conditions on the initial datum for there to exist lower and upper solsutions. We assume throughout that the initial datum, $\varphi(x)$, satisfies the conditions (1.3). We first wish to prove that there exist positive constants θ and r such that if $x_0>\theta$ then for some $\bar{x}\in(x_0-r,x_0)$ the vertical line $l_1(t)=\bar{x}$ is a lower solution on $\mathbb{R}^+$. The proof of this result is broken up into a few lemmas.

LEMMA 2.4. *Let* $\Phi(x_0)(t)=\int_0^t d\tau\int_{-x_0}^{x_0}K(x_0-\xi,t_0-\tau)\,d\xi$ *and fix* $\varepsilon\in(0,\frac{1}{2}-a)$. *There exists a positive constant* $\theta(\varepsilon)$ *such that if* $x_0>\theta(\varepsilon)$, *then* $\Phi(x_0)(t)+\Phi(x_0)'(t)\geq a+\varepsilon$ *in* $\mathbb{R}^+$.

Proof. Let

$$a_\varepsilon=a+\varepsilon,\qquad \delta=\min\left(\frac{1/2-a_\varepsilon}{2t_0},\frac{1/2-a_\varepsilon}{1+t_0}\right),$$

$$t_0=-\log\left(\frac{1}{2}-a_\varepsilon\right),\qquad \theta(\varepsilon)=\max\left(1,2t_0\log\frac{2t_0^{1/2}}{\pi^{1/2}\delta}\right).$$

Assume that $x_0 \geq \theta(\varepsilon)$. The proof will be broken into two steps. First assume that $t \in (0, t_0)$. Then, using (2.2), we have

(2.3)

$$\begin{aligned}\Phi(x_0)(t)+\Phi(x_0)'(t) &= \int_0^t d\tau \int_{-x_0}^{x_0} K(x_0-\xi, t-\tau)\,d\xi + \int_{-x_0}^{x_0} K(x_0-\xi, t)\,d\xi \\ &= \int_0^t d\tau \int_{-\infty}^{x_0} K(x_0-\xi, t-\tau)\,d\xi + \int_{-\infty}^{x_0} K(x_0-\xi, t)\,d\xi \\ &\quad - \left[\int_0^t d\tau \int_{-\infty}^{-x_0} K(x_0-\xi, t-\tau)\,d\xi + \int_{-\infty}^{-x_0} K(x_0-\xi, t)\,d\xi\right] \\ &= \frac{1}{2} - \left[\int_0^t d\tau \int_{-\infty}^{-x_0} K(x_0-\xi, t-\tau)\,d\xi + \int_{-\infty}^{-x_0} K(x_0-\xi, t)\,d\xi\right].\end{aligned}$$

We now show that for $\tau \in (0, t)$

$$\int_{-\infty}^{-x_0} K(x_0-\xi, t-\tau)\,d\xi < \delta. \tag{2.4}$$

From this and (2.3) it will follow that for $t \in (0, t_0)$

$$\Phi(x_0)(t)+\Phi(x_0)'(t) \geq \frac{1}{2} - (1+t)\delta \geq \frac{1}{2} - (1+t_0)\delta \geq a_\varepsilon.$$

Now (2.4) is true because for $\tau \in [0, t)$

$$\begin{aligned}\int_{-\infty}^{-x_0} K(x_0-\xi, t-\tau)\,d\xi &= \frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}} \int_{-\infty}^{-x_0} \exp\left(-(x_0-\xi)^2/4(t-\tau)\right) d\xi \\ &\leq \frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}} \int_{-\infty}^{-x_0} \exp\left(-(x_0-\xi)/4(t-\tau)\right) d\xi.\end{aligned}$$

The last inequality is true because $x_0 > \theta(\varepsilon) \geq 1$. Therefore,

$$\begin{aligned}\int_{-\infty}^{x_0} K(x_0-\xi, t-\tau)\,d\xi &\leq \frac{2(t-\tau)^{1/2}}{\pi^{1/2}} e^{-x_0/2(t-\tau)} e^{-(t-\tau)} \\ &\leq \frac{2t^{1/2}}{\pi^{1/2}} e^{-x_0/2t} \leq \frac{2t_0^{1/2}}{\pi^{1/2}} e^{-\theta(\varepsilon)/2t_0} \leq \delta.\end{aligned}$$

Now assume that $t \geq t_0$. Then

$$\begin{aligned}\Phi(x_0)(t)+\Phi(x_0)'(t) &\geq \Phi(x_0)(t_0) \\ &= \int_0^{t_0} d\tau \int_{-\infty}^{x_0} K(x_0-\xi, t_0-\tau)\,d\xi - \int_0^{t_0} d\tau \int_{-\infty}^{-x_0} K(x_0-\xi, t_0-\tau)\,d\xi.\end{aligned}$$

Since

$$\int_0^{t_0} d\tau \int_{-\infty}^{x_0} K(x_0-\xi, t_0-\tau)\,d\xi = \frac{1-e^{-t_0}}{2},$$

we conclude from (2.4) that

$$\Phi(x_0)(t)+\Phi(x_0)'(t)\geq\frac{1-e^{-t_0}}{2}-\delta t_0\geq a_\varepsilon. \qquad \square$$

LEMMA 2.5. *Fix* $\varepsilon\in(0,\frac{1}{2}-a)$ *and let* $\theta=\theta(\varepsilon)$. *Let* $\theta_1=\theta+(4a/\varepsilon)^{1/2}$ *and*

$$h_\varepsilon(x)=\begin{cases} a & \text{for } |x|\leq\theta,\\ 0 & \text{for } |x|\geq\theta_1,\\ a-\dfrac{\varepsilon}{4}(x-\theta)^2 & \text{for } x\in(\theta,\theta_1),\\ a-\dfrac{\varepsilon}{4}(x+\theta)^2 & \text{for } x\in(-\theta_1,-\theta). \end{cases}$$

Assume that:

a) $\varphi(x)>h_\varepsilon(x)$ *for* $|x|\leq x_1$,

b) $\varphi(x)\geq h_\varepsilon(x)=0$ *for* $|x|>x_1$.

Then there exists $\bar{x}\in(\theta,\theta_1)$ *such that the line* $l_1(t)=\bar{x}$ *is a lower solution in* $\mathbb{R}^+$.

Proof. Because of our assumptions on $\varphi(x)$ there exists a function $\varphi_1(x)$ such that

(2.5)
(a) $\varphi_1(x)\in C^\infty(-\infty,\infty)$,
(b) $h_\varepsilon(x)<\varphi_1(x)<\varphi(x)$ for $|x|\leq x_1$,
(c) $h_\varepsilon(x)\leq\varphi_1(x)\leq\varphi(x)$ for $|x|>x_1$,
(d) $\varphi_1'(x)<0$ for $x>0$,
(e) $\varphi_1(x)=\varphi_1(-x)$ in $\mathbb{R}$,
(f) $\varphi_1(x)<a+\dfrac{\varepsilon}{2}$ in $\mathbb{R}$,
(g) $\varphi_1''(x)\geq-\dfrac{\varepsilon}{2}$ in $\mathbb{R}$.

From these assumptions it follows that $\varphi_1(\bar{x})=a$ for some unique constant $\bar{x}>\theta$. Let $\psi_1(x,t)$ be the solution of (2.1) with initial datum $\varphi_1(x)$. Since $\varphi(x)\geq\varphi_1(x)$ in $\mathbb{R}^+$ it follows from the maximum principle that $\psi(x,t)\geq\psi_1(x,t)$ in $\mathbb{R}\times\mathbb{R}^+$. We show that $a-\psi_1(\bar{x},t)\leq\Phi(\bar{x})(t)$ for $t\in\mathbb{R}$. From this it follows that $a-\psi(\bar{x},t)\leq a-\psi(\bar{x},t)\leq\Phi(\bar{x})(t)$ and hence the line $l_1(t)$ is a lower solution in $\mathbb{R}^+$.

We wish to show that $a-\psi_1(\bar{x},t)\leq\Phi(\bar{x})(t)$, or $\psi_1(\bar{x},t)\geq a-\Phi(\bar{x})(t)$ for $t\in\mathbb{R}^+$. Let $g(x,t)=\varphi_1(x)-\Phi(\bar{x})(t)$. We show, using a comparison argument, that $\psi_1(x,t)\geq g(x,t)$ in $\mathbb{R}\times\mathbb{R}^+$. Since $\varphi_1(\bar{x})=a$ this certainly implies the desired result.

In order to apply the maximum principle note that

$$g(x,0)=\varphi_1(x)=\psi_1(x,0),$$

and

$$g_t-g_{xx}+g=-[\Phi(\bar{x})(t)+\Phi(\bar{x})'(t)]+\varphi_1(x)-\varphi_1''(x)$$

$$\leq-(a+\varepsilon)+\left(a+\frac{\varepsilon}{2}\right)+\frac{\varepsilon}{2}=0=\psi_{1t}-\psi_{1xx}+\psi_1.$$

In this last calculation we used Lemma 2.4 and assumptions (2.5f, g). From the maximum principle (see [12, p. 159]) we conclude that $\psi_1(x,t)\geq g(x,t)$ in $\mathbb{R}\times\mathbb{R}^+$, and the result follows. $\square$

LEMMA 2.6. *There exist positive constants r and θ such that if the initial datum $\varphi(x)$ satisfies* (1.3) *with $x_0>\theta$, then, for some $\bar{x}\in(x_0-r,x_0)$, the line $l_1(t)=\bar{x}$ is a lower solution in $\mathbb{R}^+$.*

Proof. Choose $\varepsilon\in(0,\frac{1}{2}-a)$, $r=(4a/\varepsilon)^{1/2}$, and $\theta=\theta(\varepsilon)+r$. The result now follows from the previous lemma. □

We now prove the existence of an upper solution.

LEMMA 2.7. *There exists a linear function $l_2(t)$ such that $l_2(0)=x_0$ and $l_2(t)$ is an upper solution on* $[0,a/2]$.

Proof. Recall the function $\psi(x,t)$ defined to be the solution of (2.1) with initial datum $\psi(x,0)=\varphi(x)$. From assumptions (1.3d) and (1.3f) it follows that there exist positive constants δ_1 and δ_2 such that $|\psi_t(x,t)|<\delta_2$ and $\psi_x(x,t)<-\delta_1$ in the region $(x_0/2,\infty)\times(0,a/2)$. Let $M=(1+\delta_2)/\delta_1$, and define $l_2(t)$ by $l_2(t)=Mt+x_0$.

In order to show that $l_2(t)$ is a supersolution in $[0,a/2]$, consider the curve $\beta(t)$ defined implicitly by the equation $\psi(\beta(t),t)=a-t$, $\beta(0)=x_0$. Note that $\beta'(t)=(-1-\psi_t(\beta(t),t))/\psi_x(\beta(t),t)<M$. Hence $\beta(t)<l_2(t)$ in $(0,a/2)$. From Lemma 2.1 it follows that for $t\in(0,a/2)$

$$\Theta(l_2)(t)>\theta(\beta)(t)=a-\psi(\beta(t),t)=t.$$

On the other hand,

$$\Phi(l_2)(t)=\int_0^t d\tau\int_{-l_2(\tau)}^{l_2(\tau)}K(l_2(t)-\xi,t-\tau)\,d\xi\le\int_0^t 1\,d\tau=t.$$

Therefore, $\Phi(l_2)(t)<\Theta(l_2)(t)$ for $t\in(0,a/2)$, which means that $l_2(t)$ is a supersolution in $[0,a/2]$. □

3. Existence and regularity of $s(t)$. Throughout this section we assume that there exist linear functions $l_1(t)$ and $l_2(t)$ which are respectively lower and upper solutions in $[0,T]$ for some positive time T. Recall that $s(t)$ is a solution of the integral equation (1.5) in $[0,T]$ if and only if

$$\Phi_{t_0}(s)(t)=\Theta_{t_0}(s)(t)$$

for $0\le t_0<t\le T$. We prove the existence of a solution of (1.5) in $[0,T]$ by constructing a sequence of continuous, piecewise linear functions $\{s_n(t)\}$ with the properties that $s_n(0)=x_0$ and, if we set $t_j=jT/n$,

$$\Theta_{t_j}(s_n)(t_{j+1})=\Phi_{t_j}(s_n)(t_{j+1})\quad\text{for } j=0,\cdots,n-1,\quad n=1,2,\cdots.$$

This sequence of functions is shown to be equicontinuous and uniformly bounded. Therefore, by the theorem of Arzela and Ascoli some subsequence of $\{s_n\}$ converges uniformly to a continuous function. This continuous function is shown to be a solution of the integral equation (1.5).

LEMMA 3.1. *For each positive integer n there exists a continuous piecewise linear function $s_n(t)$, defined in* $[0,T]$, *such that $l_1(t)\le s_n(t)\le l_2(t)$ and, if we set $t_j=jT/n$,*

$$\Phi_{t_j}(s_n)(t_{j+1})=\Theta_{t_j}(s_n)(t_{j+1}),\qquad j=0,1,\cdots,n-1.$$

Proof. Fix n. Set $s_n(0)=x_0$ and suppose that we have found points $x_0,x_1,\cdots,x_k$ such that $l_1(t_j)\le x_j\le l_2(t_j)$, $j=0,1,\cdots,k$, and, if $s_n(t)$ is the piecewise linear function

connecting the points (x_j, t_j), then

$$\Phi_{t_j}(s_n)(t_{j+1}) = \Theta_{t_j}(s_n)(t_{j+1}), \qquad j = 0, 1, \cdots, k-1.$$

For $x \in (l_1(t_{k+1}), l_2(t_{k+1}))$, let

$$\alpha(x)(t) = \begin{cases} s_n(t) & \text{for } t \le t_{nk}, \\ \text{the line segment connecting } (x_k, t_k) \text{ and } (x, t_{k+1}) & \text{for } t_k < t \le t_{k+1}. \end{cases}$$

By induction the proof of the lemma will be complete once we have proven the existence of a point x_{k+1} such that $l_1(t_{k+1}) \le x_{k+1} \le l_2(t_{k+1})$, and $\Phi_{t_k}(\alpha(x_{k+1}))(t_{k+1}) = \Theta_{t_k}(\alpha(x_{k+1}))(t_{k+1})$. To prove the existence of x_{k+1} we first let $x^1 = l_1(t_{k+1})$ and show that $\Phi_{t_k}(\alpha(x^1))(t_{k+1}) - \Theta_{t_k}(\alpha(x^1))(t_{k+1}) \ge 0$. We then let $x^2 = l_2(t_{k+1})$ and show that $\Phi_{t_k}(\alpha(x^2))(t_{k+1}) - \Theta_{t_k}(\alpha(x^2))(t_{k+1}) \le 0$. Since $\Phi_{t_k}(\alpha(x))(t_{k+1}) - \Theta_{t_k}(\alpha(x))(t_{k+1})$ is a continuous function of x it will then follow that there must exist a point $x_{k+1} \in [x^1, x^2]$ such that $\Phi_{t_k}(\alpha(x_{k+1}))(t_{k+1}) - \Theta_{t_k}(\alpha(x_{k+1}))(t_{k+1}) = 0$.

Note that $\alpha(x^1)(t) \ge l_1(t)$ for $t \in (0, t_{k+1})$. From Lemma 2.2 it follows that $\Theta(\alpha(x^1))(t_{k+1}) \ge \Theta(l_1)(t_{k+1})$. Therefore, since $l_1(t)$ is a lower solution, $\Phi(\alpha(x^1))(t_{k+1}) - \Theta(\alpha(x^1))(t_{k+1}) \ge \Phi(l_1)(t_{k+1}) - \Theta(l_1)(t_{k+1}) \ge 0$. Since $\alpha(x^1)(t) = s_n(t)$ for $t \in (0, t_k)$ it follows that $\Phi(\alpha(x^1))(t_k) - \Theta(\alpha(x^1))(t_k) = \Phi(s_n)(t_k) - \Theta(s_n)(t_n) = 0$. Hence,

$$\begin{aligned} &\Phi_{t_k}(\alpha(x^1))(t_{k+1}) - \Theta_{t_k}(\alpha(x^1))(t_{k+1}) \\ &\qquad = \left[\Phi(\alpha(x^1))(t_{k+1}) - \Theta(\alpha(x^1))(t_{k+1})\right] - \left[\Phi(\alpha(x^1))(t_k) - \Theta(\alpha(x^1))(t_k)\right] \\ &\qquad \ge 0. \end{aligned}$$

A similar argument shows that $\Phi_{t_k}(\alpha(x^2))(t_{k+1}) - \Theta_{t_k}(\alpha(x^2))(t_{k+1}) \le 0$. From our previous remarks this completes the proof of the lemma. □

In order to apply the theorem of Arzela and Ascoli to conclude that a subsequence of $\{s_n(t)\}$ converges uniformly to a continuous function we need to show that the sequence $\{s_n(t)\}$ is equicontinuous. We now prove this to be true if T is chosen sufficiently small.

LEMMA 3.2. *If T is chosen so that $e^{-l_1(T)/T} \le \frac{1}{4}$ then the sequence $\{s_n(t)\}$ is equicontinuous on $[0, T]$.*

Proof. Let B be the region bounded by $l_1(t)$, $l_2(t)$, $t = 0$ and $t = T$. From assumption (1.3d) it follows that $\psi_x(x,t) < 0$ in B. Choose δ_1 to be a positive constant such that $\psi_x(x,t) < -\delta_1$ in B. From assumption (1.3f) there exists a positive constant δ_2 such that $|\psi_t(x,t)| < \delta_2$ in B (see [6, Thm. 6, p. 65]). Let $M = \sup_{0 \le \tau < t \le T} K(l_1(t) + l_1(\tau), t - \tau)$ and $\bar{t} = \min(\delta_1/4M, T)$.

Since each function $s_n(t)$ is piecewise linear it suffices to show that the derivatives $s_n'(t)$ are uniformly bounded whenever they exist. We first find a lower bound on $s_n'(t)$ for $t \in [0, T]$ and $n = 0, 1, 2, \cdots$. In fact, suppose that p is a positive integer such that $p\bar{t} < T$. We show that $s_n'(t) \ge -2^p\delta_2/\delta_1$ for each n and $t \in (0, p\bar{t})$ such that $s_n'(t)$ is defined.

Suppose that this is not true. Then there must exist positive integers m and n such that $1 \le m \le p$, $s_n'(\hat{t}) < -2^m\delta_2/\delta_1$ for some $\hat{t} \in ((m-1)\bar{t}, m\bar{t})$, and $s_n'(t) \ge -2^{m-1}\delta_2/\delta_1$ for $t < (m-1)\bar{t}$. Since $s_n(t)$ is piecewise linear we may assume that for some integer k, $s_n'(t) \ge -2^m\delta_2/\delta_1$ for $t < t_k \equiv kT/n$, and $s_n'(t) < -2^m\delta_2/\delta_1$ for $t \in (t_k, t_{k+1})$. We show that $\Phi(s_n)'(t) - \Theta(s_n)'(t) > 0$ for $t \in (t_k, t_{k+1})$. This immediately leads to a contradiction because $\Phi(s_n)(t_k) - \Theta(s_n)(t_k) = \Phi(s_n)(t_{k+1}) - \Theta(s_n)(t_{k+1}) = 0$.

We first estimate $\Phi(s_n)'(t)$ for $t\in(t_k,t_{k+1})$. Using (2.2) it follows that:

$$\begin{aligned}\Phi(s_n)'(t)\geq&\left[\int_0^{(m-1)t_1}K(s_n(t)+s_n(\tau),t-\tau)(s_n'(\tau)+s_n'(t))\,d\tau\right.\\&\left.+\int_0^{(m-1)t_1}K(s_n(t)-s_n(\tau),t-\tau)(s_n'(\tau)-s_n'(t))\,d\tau\right]\\&+\left[\int_{(m-1)t_1}^{t}K(s_n(t)+s_n(\tau),t-\tau)(s_n'(\tau)+s_n'(t))\,d\tau\right.\\&\left.+\int_{(m-1)t_1}^{t}K(s_n(t)-s_n(\tau),t-\tau)(s_n'(\tau)-s_n'(t))\,d\tau\right]\\=&[\mathrm{I}]+[\mathrm{II}].\end{aligned}$$

We show that $[\mathrm{I}]>0$. Recall that for $\tau\in(0,(m-1)t_1)$, $s_n'(\tau)\geq-2^{m-1}\delta_2/\delta_1>s_n'(t)$. Hence

$$\begin{aligned}[\mathrm{I}]\geq\int_0^{(m-1)t_1}&\left[2s_n'(t)K(s_n(t)+s_n(\tau),t-\tau)\right.\\&\left.-\left[s_n'(t)+2^{m-1}\frac{\delta_2}{\delta_1}\right]K(s_n(t)-s_n(\tau),t-\tau)\right]d\tau.\end{aligned}$$

The right-hand side is positive if for each $\tau<(m-1)t_1$,

$$\begin{aligned}2s_n'(t)&\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}\exp\left(-\frac{(s_n(t)+s_n(\tau))^2}{4(t-\tau)}\right)\\&>\left[s_n'(t)+2^{m-1}\frac{\delta_2}{\delta_1}\right]\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}\exp\left(-\frac{(s_n(t)-s_n(\tau))^2}{4(t-\tau)}\right)\end{aligned}$$

or

$$\exp\left(-\frac{s_n(t)s_n(\tau)}{t-\tau}\right)<\frac{s_n'(t)+2^{m-1}\delta_2/\delta_1}{2s_n'(t)}.$$

This is true because

$$\exp\left(-\frac{s_n(t)s_n(\tau)}{t-\tau}\right)\leq\exp\left(-\frac{l_1(T)^2}{T}\right)<\frac{1}{4}$$

by assumption, and

$$\frac{s_n'(t)+2^{m-1}\delta_2/\delta_1}{2s_n'(t)}=\frac{1}{2}+2^{m-2}\frac{\delta_2}{s_n'(t)\delta_1}\geq\frac{1}{4}.$$

We have therefore shown that $[\mathrm{I}]>0$. On the other hand,

$$[\mathrm{II}]\geq\int_{(m-1)t_1}^{t}2s_n'(t)K(l_1(t)+l_1(\tau),t-\tau)\,d\tau\geq2s_n'(t)t_1M.$$

Therefore, $\Phi(s_n)'(t)>2Mt_1s_n'(t)$.

We now show that $\Theta(s_n)'(t)<2M\bar{t}s_n'(t)$ for $t\in(t_k,t_{k+1})$. This is true because

$$\begin{aligned}\Theta(s_n)'(t)&=-\psi_x(s_n(t),t)s_n'(t)-\psi_t(s_n(t),t)\\&<\delta_1 s_n'(t)+\delta_2=4M\bar{t}s_n'(t)+\delta_2\\&\le 4M\bar{t}s_n'(t)-2M\bar{t}s_n'(t)=2M\bar{t}s_n'(t).\end{aligned}$$

We have therefore shown that $\Phi(s_n)'(t)>\Theta(s_n)'(t)$ for $t\in(t_k,t_{k+1})$. As was mentioned earlier this leads to a contradiction. Hence, the uniform lower bound on $s_n'(t)$ follows. Using a similar argument one can obtain a uniform upper bound on $s_n'(t)$. In fact, if P is chosen so that $P\bar{t}<T$ then one can show that $s_n'(t)\le 2^P((1+\delta_2)/\delta_1)$ for each n and $t\in(0,P\bar{t})$ such that $s_n'(t)$ is defined (see [13] for details). From our previous remarks this concludes the proof of the lemma. □

Since the sequence $\{s_n(t)\}$ is equicontinuous, and uniformly bounded by the lower and upper solutions $l_1(t)$ and $l_2(t)$ on $[0,T]$, the theorem of Arzela and Ascoli guarantees that a subsequence, $\{s_{nk}(t)\}$, converges uniformly on $[0,T]$ to a uniformly Lipschitz function $s(t)$. To simplify notation we write $\{s_{nk}(t)\}=\{s_n(t)\}$.

LEMMA 3.4. *$s(t)$ is a solution of the integral equation* (1.5) *in* $[0,T]$.

Proof. Let ε be an arbitrary positive constant and choose $t_0\in[0,T]$. We show that $|\Phi(s)(t_0)-\Theta(s)(t_0)|<\varepsilon$ be estimating, for sufficiently large n, each term of the inequality

$$\begin{aligned}|\Phi(s)(t_0)-\Theta(s)(t_0)|\le{}&|\Phi(s)(t_0)-\Phi(s_n)(t_k)|+|\Phi(s_n)(t_k)-\Theta(s_n)(t_k)|\\&+|\Theta(s_n)(t_k)-\Theta(s)(t_0)|.\end{aligned}$$

Here k is chosen so that $t_0\in(t_k,t_{k+1})$.

It follows from the construction of $s_n(t)$ that $|\Phi(s_n)(t_k)-\Theta(s_n)(t_k)|=0$. Furthermore, because the function $\psi(x,t)$ is uniformly continuous and the sequence of functions $\{s_n(t)\}$ are uniformly Lipschitz continuous, it follows that $|\Theta(s_n)(t_k)-\Theta(s)(t_0)|<\varepsilon/2$ for n sufficiently large. It remains to show that $|\Phi(s)(t_0)-\Phi(s_n)(t_0)|<\varepsilon/2$ for n sufficiently large. Setting $\lambda=s_n(t_k)-s(t_0)$, this is true because

$$\begin{aligned}|\Phi(s)(t_0)-\Phi(s_n)(t_k)|={}&\Big|\int_0^{t_0}d\tau\int_{-s(\tau)}^{s(\tau)}K(s(t_0)-\xi,t_0-\tau)\,d\xi\\&-\int_0^{t_k}d\tau\int_{-s_n(\tau)}^{s_n(\tau)}K(s_n(t_k)-\xi,t_k-\tau)\,d\xi\Big|\\={}&\Big|\int_{t_k-t_0}^{0}d\tau\int_{-s(\tau+t_0-t_k)+\lambda}^{s_n(\tau+t_0-t_k)+\lambda}K(s_n(t_k)-\xi,t_k-\tau)\,d\xi\\&+\int_0^{t_k}d\tau\int_{-s(\tau+t_0-t_k)+\lambda}^{-s_n(\tau)}K(s_n(t_k)-\xi,t_k-\tau)\,d\xi\\&+\int_0^{t_k}d\tau\int_{s_n(\tau)}^{s(\tau+t_0-t_k)+\lambda}K(s_n(t_k)-\xi,t_k-\tau)\,d\xi\Big|\\\le{}&|t_k-t_0|+4\sup_{0\le\tau\le t_k}|s(\tau+t_0-t_k)-s_n(\tau)|\int_0^{t_k}\frac{d\tau}{2\pi^{1/2}(t_k-\tau)^{1/2}}\\\le{}&\frac{\varepsilon}{2}\end{aligned}$$

if n is sufficiently large. In the last inequality we used the fact that $s(t)$ is a Lipschitz continuous function and $|t_k - t_0| < 1/n$. □

Note that the results of [14] now imply that $s(t) \in C^1(0,T)$, and $s'(t)$ is a locally Lipschitz continuous function. In [14] it is shown that this is true for $t \in (0,\delta)$ for some $\delta > 0$. We may choose $\delta = T$ as long as we know that $s(t) > 0$ and $s'(t)$ is bounded in $(0,T)$. Both of these conditions are satisfied. It is true that $s(t) > 0$ in $(0,T)$ from the assumption that there exists a lower solution $l_1(t)$. Furthermore, since it has been proven that $s(t)$ is uniformly Lipschitz continuous, it follows that $s'(t)$ is bounded wherever it is defined.

We have now shown that if T is chosen so that there exist linear functions $l_1(t)$ and $l_2(t)$ which are, respectively, lower and upper solutions in $[0,T]$, and $e^{-l_1(T)/T} \leq \frac{1}{4}$, then there exists a smooth function $s(t)$ which satisfies (1.5) in $[0,T]$. Combining this result with Lemmas 2.6 and 2.7 it follows that there exist positive constants r and θ (as in Lemma 2.6) such that if $\varphi(x)$ satisfies (1.3) with $x_0 > \theta$ then $s(t)$ is a smooth function defined in all of $\mathbb{R}^+$. Furthermore, $s(t) > x_0 - r$ in $\mathbb{R}^+$.

4. Uniqueness. The following theorem demonstrates that the solution of (1.5) is unique among uniformly Lipschitz functions.

THEOREM 4.1. *Suppose that $\alpha(t)$ and $\beta(t)$ are respectively lower and upper solutions in $[0,T]$. Then $\alpha(t) \leq \beta(t)$ in $[0,T]$.*

Proof. Note that we must have $\alpha(0) \leq x_0 \leq \beta(0)$. If, for example, $\alpha(0) > x_0$, then $\psi(\alpha(0),0) < a$. It follows there must exist some time, t_0, such that $\psi(\alpha(t),t) < a - t$ for $t \in (0,t_0)$. Therefore, $\Theta(\alpha)(t) = a - \psi(\alpha(t),t) > t$ for $t \in (0,t_0)$. On the other hand,

$$\Phi(\alpha)(t) = \int_0^t d\tau \int_{-s(\tau)}^{s(\tau)} K(\alpha(t) - \xi, t - \tau)\, d\xi \leq \int_0^t 1\, d\tau = t$$

for all $t \in \mathbb{R}^+$. Hence, $\Theta(\alpha)(t) > \Phi(\alpha)(t)$ in $(0,t_0)$, which contradicts the assumption that $\alpha(t)$ is a subsolution. A similar argument shows that it is impossible for $\beta(0) < x_0$.

If $\alpha(0) < \beta(0)$, then we must have $\alpha(t) < \beta(t)$ in $(0,T)$. If not, we let $t_0 = \inf\{t : \alpha(t) \geq \beta(t)\}$. Then $\alpha(t_0) = \beta(t_0)$ and $\alpha(t) < \beta(t)$ in $(0,t_0)$. Lemma 2.1 now implies that $\Theta(\alpha)(t_0) > \Theta(\beta)(t_0)$, while Lemma 2.2 implies that $\Phi(\alpha)(t_0) < \Phi(\beta)(t_0)$. Since $\alpha(t)$ is a lower solution and $\beta(t)$ an upper solution, we now have

$$\Theta(\alpha)(t_0) \leq \Phi(\alpha)(t_0) < \Phi(\beta)(t_0) \leq \Theta(\beta)(t_0) < \Theta(\alpha)(t_0).$$

This is an obvious contradiction.

Throughout the rest of the proof we assume that $\alpha(0) = \beta(0) = x_0$.

Suppose the lemma is not true, and let $t_0 = \inf\{t | \alpha(t) > \beta(t)\}$. Then, $\alpha(t) = \beta(t)$ for $t \in [0,t_0]$. This is because, if $\alpha(t) < \beta(t)$ for some $t \in [0,t_0]$, it would follow from Lemmas 2.1 and 2.2 that

$$\Phi(\alpha)(t_0) < \Phi(\beta)(t_0) \leq \Theta(\beta)(t_0) = \Theta(\alpha)(t_0).$$

This, however, contradicts the assumption that $\alpha(t)$ is a lower solution.

We prove the lemma by showing that there exists some $t > t_0$ such that $\alpha(t) > \beta(t)$ and $\Phi(\alpha)(t) < \Phi(\beta)(t)$. This leads to a contradiction for the following reason. Since $\alpha(t) > \beta(t)$, and $\alpha(0) = \beta(0)$, it follows from Lemma 2.1 that $\theta(\alpha)(t) > \Theta(\beta)(t)$. If it is also true that $\Phi(\alpha)(t) < \Phi(\beta)(t)$, then, since $\beta(t)$ is an upper solution, $\Phi(\alpha)(t) < \Phi(\beta)(t) \leq \Theta(\beta)(t) < \Theta(\alpha)(t)$. This, however, contradicts the assumption that $\alpha(t)$ is a lower solution on $[0,T]$.

For $t>t_0$, let $\varepsilon(t)=\alpha(t)-\beta(t)$. Choose $\bar{t}>t_0$ such that $\varepsilon(\bar{t})>0$ and $\varepsilon(t)<\varepsilon(\bar{t})$ in $(0,\bar{t})$. Then,

$$\begin{aligned}\Phi(\beta)(\bar{t})&=\int_0^{\bar{t}}d\tau\int_{-\beta(\tau)}^{\beta(\tau)}K(\beta(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\\&=\int_0^{\bar{t}}d\tau\int_{-\beta(\tau)+\varepsilon(\bar{t})}^{\beta(\tau)+\varepsilon(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\\&=\Phi(\alpha)(\bar{t})+\left[\int_0^{t_0}d\tau\int_{\alpha(\tau)}^{\beta(\tau)+\varepsilon(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\right.\\&\qquad\left.-\int_0^{t_0}d\tau\int_{-\alpha(\tau)}^{-\beta(\tau)+\varepsilon(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\right]\\&\quad+\left[\int_{t_0}^{\bar{t}}d\tau\int_{\alpha(\tau)}^{\beta(\tau)+\varepsilon(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\right.\\&\qquad\left.-\int_{t_0}^{\bar{t}}d\tau\int_{-\alpha(\tau)}^{-\beta(\tau)+\varepsilon(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\right]\\&=\Phi(\alpha)(\bar{t})+[\mathrm{I}]+[\mathrm{II}].\end{aligned}$$

Recall that we wish to choose $\bar{t}$ so that $\Phi(\beta)(\bar{t})>\Phi(\alpha)(\bar{t})$. Note that $[\mathrm{I}]>0$. This is because, if $(\xi,\tau)\in(0,\varepsilon(\bar{t}))\times(0,t_0)$, then $|\alpha(\bar{t})-(\beta(\tau)+\xi)|<|\alpha(\bar{t})+\beta(\tau)-\xi|$, and, therefore, $K(\alpha(\bar{t})-(\beta(\tau)+\xi),\bar{t}-\tau)>K(\alpha(\bar{t})+\beta(\tau)-\xi,\bar{t}-\tau)$.

To complete the proof of the lemma it remains to choose $\bar{t}$ so that $[\mathrm{II}]>0$. We rewrite [II] as

$$[\mathrm{II}]=\int_{A_1(\bar{t})}\int K(\alpha(\bar{t})-\xi,\bar{t}-\tau)\,d\xi\,d\tau-\int_{A_2(\bar{t})}\int K(\alpha(\bar{t})-\xi,t-\tau)\,d\xi\,d\tau$$

where

$$A_1(\bar{t})=\{(\xi,\tau):t_0\le\tau\le\bar{t},\ \alpha(\tau)\le\xi\le\beta(\tau)+\varepsilon(\bar{t})\},$$

$$A_2(\bar{t})=\{(\xi,\tau):t_0\le\tau\le\bar{t},\ -\alpha(\tau)\le\xi\le-\beta(\tau)+\varepsilon(\bar{t})\}.$$

Let

$$\lambda_1(\bar{t})=\inf_{(\xi,\tau)\in A_1(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau),$$

$$\lambda_2(\bar{t})=\sup_{(\xi,\tau)\in A_2(\bar{t})}K(\alpha(\bar{t})-\xi,\bar{t}-\tau).$$

Then $[\mathrm{II}]\ge\lambda_1(\bar{t})\mu(A_1(\bar{t}))-\lambda_2(\bar{t})\mu(A_2(\bar{t}))$ where μ is Lebesgue measure on $\mathbb{R}^2$.

We now show that $\lim_{t\downarrow t_0}\lambda_1(t)=\infty$ and $\lim_{t\downarrow t_0}\lambda_2(t)=0$. The first limit follows because both $\alpha(t)$ and $\beta(t)$ are uniformly Lipschitz continuous. That is, there exists a constant L such that if $t>t_0$ and $\varepsilon(t)>0$, then $|\alpha(t)-\xi|\le L(t-\tau)$ for all $(\xi,\tau)\in A_1(t)$.

Therefore, if $(\xi,\tau)\in A_1(t)$, then

$$K(\alpha(t)-\xi,t-\tau)=\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}\exp\left(-\frac{(\alpha(t)-\xi)^2}{4(t-\tau)}\right)$$

$$\geq\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}\exp\left(-\frac{L^2}{4}(t-\tau)\right)$$

$$\geq\frac{e^{(t-t_0)}}{2\pi^{1/2}(t-t_0)^{1/2}}\exp\left(-\frac{L^2}{4}(t-t_0)\right).$$

Hence $\lambda_1(t)=\inf_{(\xi,\tau)\in A_1(t)}K(\alpha(t)-\xi,t-\tau)\to\infty$ as $t\downarrow t_0$.

On the other hand, $\lambda_2(t)\to 0$ as $t\downarrow t_0$ for the following reason. If $(\xi,\tau)\in A_2(\tau)$, then $\xi<0$. Hence, $\alpha(t)-\xi>\alpha(t)$. Therefore, for $(\xi,\tau)\in A_2(t)$,

$$K(\alpha(t)-\xi,t-\tau)\leq\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}\exp\left(-\frac{\alpha(t)^2}{4(t-\tau)}\right).$$

From this it follows that $\lambda_2(t)=\sup_{(\xi,\tau)\in A_2(t)}K(\alpha(t)-\xi,t-\tau)\to 0$ as $t\downarrow t_0$.

Now choose $t_1>t_0$ so that $\varepsilon(t_1)>0$, $\varepsilon(t)<\varepsilon(t_1)$ for $t\in(t_0,t_1)$, and $\lambda_1(t)>4\lambda_2(t)$ for $t\in(t_0,t_1)$. Let $h(t)=\beta(t)+(t/t_1)\varepsilon(t_1)$. We consider two cases.

Case 1. Suppose there exists $\bar{t}\in(t_0,t_1)$ such that $\alpha(t)\leq h(t)$ for all $t\leq\bar{t}$, and $\alpha(\bar{t})=h(\bar{t})$. Let $B(\bar{t})=\{(x,t):t_0\leq t\leq\bar{t};\ h(t)\leq x\leq\beta(t)+\varepsilon(\bar{t})\}$. Then $B(\bar{t})\subset A_1(\bar{t})$, and $\mu(B(\bar{t}))=\frac{1}{2}\varepsilon(\bar{t})\bar{t}$. Therefore, $\mu(A_1(\bar{t}))\geq\frac{1}{2}\varepsilon(\bar{t})\bar{t}$. On the other hand, $\mu(A_2(\bar{t}))\leq 2\varepsilon(\bar{t})\bar{t}$. It now follows that

$$[\mathrm{II}]>\lambda_1(\bar{t})\mu(A_1(\bar{t}))-\lambda_2(\bar{t})\mu(A_2(\bar{t}))$$

$$\geq 4\lambda_2(\bar{t})\frac{1}{2}\varepsilon(\bar{t})\bar{t}-\lambda_2(\bar{t})2\varepsilon(\bar{t})\bar{t}=0.$$

Case 2. Suppose there exists a sequence $\{t_k\}$ such that $t_k\downarrow t_0$, $\alpha(t_k)>h(t_k)$, and $\varepsilon(t)<\varepsilon(t_k)$ for $t<t_k$.

Let L be a uniform Lipschitz constant for both $\alpha(t)$ and $\beta(t)$. Choose k so that $\lambda_1(t_k)>(8Lt_1/\varepsilon(t_1))\lambda_2(t_2)$.

Let

$$\delta_1(t)=-L(t-t_0)+\alpha(t_0)\quad\text{for } t>t_0,$$

$$\delta_2(t)=L(t-t_0)+\alpha(t_0)\quad\text{for } t>t_0,$$

$$Q=\{(x,t)|\delta_2(t)\leq x\leq\delta_1(t)+\varepsilon(t_k),\ t_0\leq t\}.$$

Then $A_1(t_k)\supset Q$, and $\mu(Q)=[\varepsilon(t_k)]^2/4L$. Therefore, $\mu(A_1(t_k))>[\varepsilon(t_k)]^2/4L$. As before, $\mu(A_2(t_k))<2\varepsilon(t_k)t_k$. Note that $\varepsilon(t_k)\geq(t_k/t_1)\varepsilon(t_1)$. This is because $\alpha(t_k)>h(t_k)=\beta(t_k)+(t_k/t_1)\varepsilon(t_1)$, and hence, $\varepsilon(t_k)=\alpha(t_k)-\beta(t_k)>(t_k/t_1)\varepsilon(t_1)$.

Letting $\bar{t}=t_k$ it now follows that

$$\begin{aligned}[\text{II}] &> \lambda_1(t_k)\mu(A_1(t_k))-\lambda_2(t_k)\mu(A_2(t_k)) \\ &> \frac{8Lt_1}{\varepsilon(t_1)}\lambda_2(t_k)\frac{1}{4L}[\varepsilon(t_k)]^2-\lambda_2(t_k)2\varepsilon(t_k)t_k \\ &= \frac{t_1}{\varepsilon(t_1)}\lambda_2(t_k)[\varepsilon(t_k)]^2-\varepsilon(t_k)\lambda_2(t_k)t_k \\ &\geq \frac{t_1}{\varepsilon(t_1)}\lambda_2(t_k)\varepsilon(t_k)\frac{t_k}{t_1}\varepsilon(t_1)-2\varepsilon(t_k)\lambda_2(t_k)t_k \\ &= 2\varepsilon(t_k)\lambda_2(t_k)t_k-2\varepsilon(t_k)\lambda_2(t_k)t_k=0.\end{aligned}$$

Therefore, [II]>0, and the proof of the lemma is complete. □

5. Asymptotic behavior of $s(t)$. In this section we show that if $\varphi(x)$ is super-threshold, then $\lim_{t\to\infty}(s(t)-c^*t)$ exists for some constant c^* which depends only on the parameter a. One way to think about c^* is that it is the speed of the unique (up to translation) traveling wave solution of (1.1) satisfying $\lim_{z\to-\infty}U(z)=1$ and $\lim_{z\to\infty}U(z)=0$ (see [1]). Recall that a traveling wave solution of (1.1) is a solution of the form $u(x,t)=U(z)$ where $z=x-ct$. In this paper it will be useful to think about c^* in a different fashion which we now describe.

Suppose that $u(x,t)=U(z)$, $z=x-c^*t$, is the unique traveling wave solution of (1.1) satisfying $\lim_{z\to-\infty}U(z)=1$ and $\lim_{z\to+\infty}U(z)=0$. Let $\sigma(t)$ be defined implicitly as $u(\sigma(t),t)=a$. Since the translate of a traveling wave is also a traveling wave we may assume that $\sigma(0)=0$. Then $\sigma(t)$ is given explicitly as $\sigma(t)=c^*t$. A derivation similar to that given for (1.5) shows that $\sigma(t)$ must satisfy the integral equation:

$$a-\int_{-\infty}^{\infty}K(\sigma(t)-\xi,t)U(\xi)\,d\xi=\int_0^t\int_{-\infty}^{\sigma(\tau)}K(\sigma(t)-\xi,t-\tau)\,d\xi\,d\tau.$$

Using the change in variables $\eta=\tau-t$, $\zeta=\xi-c^*t$ in the integral on the right-hand side of this equation, we find that

$$a-\int_{-\infty}^{\infty}K(c^*t-\xi,t)U(\xi)\,d\xi=\int_{-t}^{0}\int_{-\infty}^{c^*\eta}K(-\zeta,-\eta)\,d\zeta\,d\eta \tag{5.1}$$

for each $t\in\mathbb{R}^+$. Letting $t\to\infty$ in (2.1) we find that c^* must satisfy the equation:

$$a=\int_{-\infty}^{0}\int_{-\infty}^{c^*\tau}K(-\xi,-\tau)\,d\xi\,d\tau. \tag{5.2}$$

To see that there must exist a unique solution, c^*, of (5.2) we let $h(c)$ be the function defined by

$$h(c)=\int_{-\infty}^{0}\int_{-\infty}^{c\tau}K(-\xi,-\tau)\,d\xi\,d\tau. \tag{5.3}$$

Note that $h(0)=1/2$, $h'(c)<0$ for $c\in(0,1/2)$, and $\lim_{c\uparrow\infty}h(c)=0$. Since $a\in(0,1/2)$ there must exist a unique solution of (5.2).

We now show that for x_0 sufficiently large, both $\liminf_{t\to\infty}(s(t)-c^*t)$ and $\limsup_{t\to\infty}(s(t)-c^*t)$ exist. This is done by constructing continuous functions $\alpha(t)$ and $\beta(t)$ which satisfy $\alpha(t)<s(t)<\beta(t)$, and both $\lim_{t\to\infty}(\alpha(t)-c^*t)$ and $\lim_{t\to\infty}(\beta(t)-c^*t)$ exist. The construction of $\alpha(t)$ goes as follows.

Let $M=-h'(c^*)$. Then $M>0$, and, since $h(c^*)=a$, there exists a positive constant ε such that $\varepsilon<c^*/2$, and if $|\delta|<\varepsilon$, then $h(c^*-\delta)>a+\delta M/2$.

For $(x,t)\in\mathbb{R}\times\mathbb{R}^+$, let

$$B(t)=\int_{-\infty}^{0}\int_{-\infty}^{\infty}K(x-\xi,t-\tau)\,d\xi\,d\tau=e^{-t} \tag{5.4}$$

and

$$g(x,t)=\int_{0}^{t}\int_{-\infty}^{0}K(x-\xi,t-\tau)\,d\xi\,d\tau. \tag{5.5}$$

Note that if $(x,t)\in(1,\infty)\times\mathbb{R}^+$, then

$$g(x,t)<2e^{-x}. \tag{5.6}$$

This is because, if $(x,t)\in(1,\infty)\times\mathbb{R}^+$, then

$$\begin{aligned} g(x,t)&=\int_{0}^{t}\int_{-\infty}^{0}\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}e^{-(x-\xi)^2/4(t-\tau)}\,d\xi\,d\tau\\ &<\int_{0}^{t}\int_{-\infty}^{0}\frac{e^{-(t-\tau)}}{2\pi^{1/2}(t-\tau)^{1/2}}e^{-(x-\xi)/4(t-\tau)}\,d\xi\,d\tau\\ &=\frac{2}{\pi^{1/2}}e^{-x}\int_{0}^{t}(t-\tau)^{1/2}e^{-(t-\tau)}\,d\tau<2e^{-x}. \end{aligned}$$

Before we construct $\alpha(t)$, it is necessary to define a few more constants. Let $r=\min\{c^*-\varepsilon,1\}$. Choose N so large that $(3/M)e^{-rN}<\varepsilon$, and let

$$\lambda=c^*N+\frac{6}{Mr}e^{-rN}. \tag{5.7}$$

Finally, choose θ so large that if $x_0>\theta$, then $s(t)>\lambda$ in $\mathbb{R}^+$. This is possible because of Lemma 2.6.

Let

$$\alpha(t)=\begin{cases}\lambda & \text{for } t\in[0,N),\\ c^*t+\dfrac{3}{Mr}[e^{-rt}+e^{-rN}] & \text{for } t\geq N.\end{cases}$$

Note that $\alpha(t)$ is a continuous function and

$$\alpha'(t)=c^*-\frac{3}{M}e^{-rt} \tag{5.8}$$

for $t>N$. Hence, $\alpha'(t)$ is increasing for $t>N$, and $\lim_{t\to\infty}\alpha'(t)=c^*$.

We now show that $\alpha(t)<s(t)$ in $\mathbb{R}^+$. Clearly this is true for $t\in[0,N]$. To prove that $\alpha(t)<s(t)$ for $t>N$ we show that $\Phi(\alpha)(t)>a$ for $t>N$. This will imply the desired result for the following reason. Suppose that $\alpha(T)=s(T)$ for some $T>N$. We assume that $\alpha(t)<s(t)$ for $t<T$. Then, from Lemma 2.2, $\Phi(s)(t)>\Phi(\alpha)(T)>a$. However, $\Theta(s)(T)=a-\psi(s(T),T)<a$. Since $\Phi(s)(T)=\Theta(s)(T)$ this is impossible.

So suppose that $T>N$. We wish to show that $\Phi(\alpha)(T)>a$. Let $l(t)$ be the line tangent to $\alpha(t)$ at $t=T$. That is, $l(t)=\alpha'(T)(t-T)+\alpha(T)$. Since $\alpha''(t)>0$ for $t>N$, it follows that $\alpha(t)>l(t)$ in $(0,T)$. It follows from Lemma 2.2 that $\Phi(\alpha)(T)>\Phi(l)(T)$. We prove that $\Phi(l)(T)>a$.

Note that $l(0)>0$. This is because

$$l(0)=-\alpha'(T)T+\alpha(T)$$
$$=-\left[c^*-\frac{3}{M}e^{-rT}\right]T+c^*T+\frac{3}{Mr}[e^{-rT}+e^{-rN}]>0.$$

This implies that

$$\Phi(l)(T)=\int_{-\infty}^{T}\int_{-\infty}^{l(\tau)}K(l(T)-\xi,T-\tau)\,d\xi\,d\tau$$
$$-\int_{-\infty}^{0}\int_{-\infty}^{l(\tau)}K(l(T)-\xi,T-\tau)\,d\xi\,d\tau$$
$$-\int_{0}^{T}\int_{-\infty}^{-l(\tau)}K(l(T)-\xi,T-\tau)\,d\xi\,d\tau$$
$$\geq h(\alpha'(T))-B(T)-g(l(T),T).$$

Now, $h(\alpha'(T))=h(c^*-(3/M)e^{-rT})$. Since $(3/M)e^{-rN}<\varepsilon$, and $T>N$, it follows that $h(\alpha'(T))>a+3e^{-rT}$. On the other hand, $B(T)=e^{-T}\leq e^{-rT}$. To estimate $g(\alpha(T),T)$, note that, since $l(0)>0$, it follows that $\alpha(T)>\alpha'(T)T$. Hence, (5.6) implies that

$$g(l(T),T)<2e^{-\alpha'(T)T}\leq 2e^{-(c^*-\varepsilon)T}\leq 2e^{-rT}.$$

These comments prove that $\Phi(l)(T)>a$ and, hence, $\Phi(\alpha)(T)>a$. This completes the construction of $\alpha(t)$, and the proof that $\alpha(t)$ has the desired properties.

We now construct $\beta(t)$. Let $t_n=-\log(a/2n)$, $n=1,2,\cdots$, and choose c_n so that $h(c_n)=a-a/2n$. Note that $c_n\downarrow c^*$ as $n\to\infty$. Let $\lambda_1=2\sup_{0<t<t_1}s(t)$. For $t\in[0,t_1]$ let $\beta(t)=c_1t+\lambda_1$, and for $t>t_1$ let $\beta(t)$ to be the continuous, piecewise-linear function defined by $\beta'(t_n)=c_n$ for $t\in(t_n,t_{n+1})$. Clearly $\beta(t)$ is a well defined function which satisfies $\beta'(t)\to c^*$ as $t\to\infty$. It remains to prove that $s(t)<\beta(t)$ in $\mathbb{R}^+$.

Certainly $s(t)<\beta(t)$ in $[0,t_1]$. Suppose there exists $T>t_1$ such that $s(T)=\beta(T)$ and $s(t)<\beta(t)$ for $t<T$. Assume that $T\in[t_n,t_{n+1}]$. We show that this must imply that

$$\Theta(s)(T)>a-\frac{a}{2n}>\Phi(s)(T),$$

which contradicts the fact that $s(t)$ is a solution of (1.5).

First of all, note that $\psi(s(T),T)<e^{-T}$. In fact, $\psi(x,t)<e^{-t}$ for all $(x,t)\in\mathbb{R}^+\times\mathbb{R}^+$. This follows from the maximum principle applied to (2.1) and our assumption that $\psi(x,0)\in[0,1]$. Since $T\in[t_n,t_{n+1})$ we have that $\psi(s(T),T)<e^{-t_n}=a/2n$. It now follows from the definition of Θ that $\Theta(s)(T)>a-a/2n$.

It remains to prove that $\Phi(s)(T)<a-a/2n$. Let $l(t)$ be the line defined by

$$l(t)=c_n(t-T)+s(T).$$

Then $s(t)\leq\beta(t)\leq l(t)$ in $(0,T)$. From Lemma 2.2, it follows that $\Phi(s)(T)\leq\Phi(l)(T)$. However,

$$\Phi(l)(T)<\int_{-\infty}^{T}\int_{-\infty}^{l(\tau)}K(l(T)-\xi,T-\tau)\,d\xi\,d\tau=h(c_n).$$

Since $h(c_n)=a-a/2n$, we obtain the desired contradiction.

Before continuing with the proof that $\lim_{t\to\infty}(s(t)-c^*t)$ exists we introduce some notation which will be used throughout the rest of this paper.

Let $\lambda=\limsup_{t\to\infty}(s(t)-c^*)$. Choose $\{t_n\}$, $n=1,2,\cdots$, so that

$$s(t_n)>c^*t+\lambda-1/n, \tag{5.8a}$$

$$s(t)<c^*t+\lambda+1/n \quad \text{for } t>t_n. \tag{5.8b}$$

Let

$$l_n(t)=c^*(t-t_n)+s(t_n),$$
$$J_n=\{t<t_n : l_n(t)<s(t)\},$$
$$H_n=\{t<t_n : s(t)\le l_n(t)\},$$
$$A_n=\int_{J_n}\int_{l_n(\tau)}^{s(\tau)} K(s(t_n)-\xi,t_n-\tau)\,d\xi\,d\tau,$$
$$B_n=\int_{H_n}\int_{s(\tau)}^{l_n(\tau)} K(s(t_n)-\xi,t_n-\tau)\,d\xi\,d\tau.$$

Note that (5.8) implies that $s(t)<l_n(t)+2/n$ if $t>t_n$.

The next couple of lemmas give us some sort of estimate of how much the curve $s(t)$ can oscillate for $t<t_n$, $n=1,2,\cdots$. They demonstrate that for n large, and $t<t_n$, the curve $s(t)$ must be very close to the line $l_n(t)$ in some sort of weighted L^1 sense.

LEMMA 5.1. *$A_n\to 0$ as $n\to\infty$.*

Proof. Fix $m<n$. Since $s(t)<l_m(t)+2/m$ for $t>t_m$, it follows that

$$A_n\le\int_{-\infty}^{t_m}\int_{-\infty}^{\infty} K(s(t_n)-\xi,t_n-\tau)\,d\xi\,d\tau+\int_{t_m}^{t_n}\int_{l_n(\tau)}^{l_m(\tau)+2/m} K(s(t_n)-\xi,t_n-\tau)\,d\xi\,d\tau$$
$$=[\mathrm{I}]+[\mathrm{II}].$$

Now,

$$[\mathrm{I}]\le\int_{-\infty}^{t_m} e^{-(t_n-\tau)}\,d\tau\le e^{-(t_n-t_m)}.$$

On the other hand,

$$[\mathrm{II}]\le\int_{-\infty}^{t_n}\int_{l_n(\tau)}^{l_n(\tau)+4/m} K(s(t_n)-\xi,t_n-\tau)\,d\xi\,d\tau$$
$$=\int_0^{4/m}\int_{-\infty}^{0} K(-\eta-c^*\tau,-\tau)\,d\tau\,d\eta\le\frac{M}{m}$$

for some constant M which does not depend on m and n. We have shown that

$$A_n\le\frac{M}{m}+e^{-(t_n-t_m)}$$

for all $m\le n$. Let $m=n/2$ if n is even and $m=(n+1)/2$ if n is odd. It follows that $A_n\le 2M/n+e^{-t_n/2}$, and the proof of the lemma is complete. □

LEMMA 5.2. $B_n \to 0$ *as* $n \to \infty$.

Proof. Note that

$$\Phi(s)(t_n) = \int_{-\infty}^{t_n} \int_{-\infty}^{l_n(\tau)} K(s(t_n)-\xi, t_n-\tau)\, d\xi\, d\tau - \int_{-\infty}^{0} \int_{-\infty}^{l_n(\tau)} K(s(t_n)-\xi, t_n-\tau)\, d\xi\, d\tau$$

$$+ \int_{0}^{t_n} \int_{l_n(\tau)}^{s(\tau)} K(s(t_n)-\xi, t_n-\tau)\, d\xi\, d\tau - \int_{0}^{t_n} \int_{-\infty}^{s(\tau)} K(s(t_n)-\xi, t_n-\tau)\, d\xi\, d\tau$$

$$\equiv [\mathrm{I}] - [\mathrm{II}] + [\mathrm{III}] - [\mathrm{IV}] < [\mathrm{I}] + [\mathrm{III}].$$

Now, $[\mathrm{I}] = h(c^*) = a$, while, $[\mathrm{III}] = A_n - B_n$. Since

$$\Phi(s)(t_n) = \Theta(s)(t_n) = a - \psi(s(t_n), t_n)$$

it follows that

$$B_n \le A_n + \psi(s(t_n), t_n).$$

Since each term on the right-hand side of this equation $\to 0$ as $n \to \infty$, the result follows.

Here we briefly outline how the proof that $\lim_{t\to\infty}(s(t) - c^* t)$ exists will be completed. For each n we construct a sequence of positive constants $\{\delta_{nk}\}$, $k = 0, 1, \cdots$, with the property that if $\delta_n = \sum_{k=0}^{\infty} \delta_{nk}$, then $\delta_n \to 0$ as $n \to \infty$. Furthermore, letting $h_n(t)$ be the piecewise continuous function defined by

$$h_n(t) = \begin{cases} s(t) & \text{for } t \le t_n, \\ l_n(t) - \sum_{j=0}^{k} \delta_{nj} & \text{for } t_n + k < t \le t_n + k + 1, \end{cases} \tag{5.9}$$

we show that $h_n(t) \le s(t)$ for each n. This implies that $l_n(t) - \delta_n < s(t)$ for each n and $t > t_n$. Since we already know that $s(t) < l_n(t) + 2/n$ for each n and $t > t_n$, this will complete the proof. In what follows we set $t_{nk} \equiv t_n + k$.

The δ_{nk} are defined inductively. Fix n and suppose we have already chosen $\delta_{n1}, \cdots, \delta_{n,k-1}$. Furthermore assume that for $t < t_{nk}$, $h_n(t) < s(t)$ where $h_n(t)$ is defined by (5.9). We show how to define δ_{nk}. It must be chosen in such a way that $h_n(t) < s(t)$ for $t \in (t_{nk}, t_{n,k+1}]$. From the definition of δ_{nk} it will be clear that if $\delta_n \equiv \sum_{k=0}^{\infty} \delta_{nk}$, then $\delta_n \to 0$ as $n \to \infty$.

LEMMA 5.3. *If* δ_{nk} *is sufficiently large then* $\Phi(h_{nk})(t) > a$ *for each* $t \in (t_{nk}, t_{n,k+1}]$.

Before proving the lemma we show that it implies that $h_n(t) < s(t)$ in $(t_{nk}, t_{n,k+1}]$. If this were not true, then there must exist some $T \in (t_{nk}, t_{n,k+1}]$ such that $h_n(T) = s(T)$, and $h_n(t) < s(t)$ for all $t < T$. From Lemma 2.2 this would imply that $\Phi(s)(T) > \Phi(h_n)(T) > a$. Since $\Theta(s)(t) < a$ for all t this is impossible.

Proof of Lemma 5.3. Assume that $t \in (t_{nk}, t_{n,k+1}]$. To simplify the notation we set $h_{nj}(t) = l_n(t) - \sum_{i=0}^{j} \delta_{ni}$ for $t \in \mathbb{R}$. Then (5.9) becomes

$$h_n(t) = \begin{cases} s(t) & \text{for } t \le t_n, \\ h_{nj}(t) & \text{for } t \in (t_{nj}, t_{n,j+1}]. \end{cases}$$

Note that

$$\begin{aligned}\Phi(h_n)(t)&=\int_{-\infty}^{t}\int_{-\infty}^{h_{nk}(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&\quad-\left(\int_{-\infty}^{0}\int_{-\infty}^{h_{nk}(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\right.\\&\qquad\left.+\int_{0}^{t}\int_{-\infty}^{-h_n(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\right)\\&\quad+\left(\int_{0}^{t}\int_{h_{nk}(\tau)}^{h_n(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\right)\\&=a-[\mathrm{I}]+[\mathrm{II}].\end{aligned}$$

Now, using (5.4) and (5.6), we have

$$[\mathrm{I}]\le B(t)+g(h_n(t),t)\le e^{-t}+2e^{-c^*t}\le 3e^{-rk}e^{-rt_n}.$$

Here we set $r=\min(1,c^*)$. Next consider [II]. Using the fact that $h_n(\tau)=h_{nk}(\tau)$ for $\tau\in(t_{nk},t)$, we rewrite [II] as

$$\begin{aligned}[\mathrm{II}]&=\int_{0}^{t_{nk}}\int_{h_{nk}(\tau)}^{h_{n,k-1}(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau-\int_{0}^{t_{nk}}\int_{h_n(\tau)}^{h_{n,k-1}(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&=[\mathrm{II}_1]-[\mathrm{II}_2].\end{aligned}$$

Note that

$$\begin{aligned}[\mathrm{II}_1]&\ge\int_{t-2}^{t-1}\int_{h_{nk}(\tau)}^{h_{n,k-1}(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&=\int_{0}^{\delta_{nk}}\int_{-2}^{-1}K(-c^*t-\eta,-\tau)\,d\tau\,d\eta\ge\delta_{nk}M_1\end{aligned}$$

for some constant M_1 which does not depend on n. On the other hand, $h_n(\tau)=s(\tau)$ for $\tau<t_n$, $h_{n,k-1}(\tau)<h_n(\tau)$ for $\tau\in(t_n,t_{n,k-1})$, and $h_{n,k-1}(\tau)<l_n(\tau)$ for $\tau<t_n$. Therefore,

$$\begin{aligned}[\mathrm{II}_2]&<\int_{0}^{t_n}\int_{h_n(\tau)}^{h_{n,k-1}(\tau)}K(h_n(t)-1,t-\tau)\,d\xi\,d\tau\\&<\int_{H_n}\int_{s(\tau)}^{l_n(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&=\int_{H_n\cap(0,t_n-1)}\int_{s(\tau)}^{l_n(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&\quad+\int_{H_n\cap(t_n-1,t_n)}\int_{s(\tau)}^{l_n(\tau)}K(h_n(t)-\xi,t-\tau)\,d\xi\,d\tau\\&=[\mathrm{II}_{21}]+[\mathrm{II}_{22}].\end{aligned}$$

To estimate $[II_{21}]$ we set $G_n=\{(\xi,\tau):\tau\in H_n\cap(0,t_{n-1}),\ s(\tau)<\xi<l_n(\tau)\}$, and $P_n(t)=\sup_{(\xi,\tau)\in G_n}K(h_n(t)-\xi,t-\tau)/K(s(t_n)-\xi,t_n-\tau)$. Then

$$[II_{21}]=\int_{G_n}\int K(s(t_n)-\xi,t_n-\tau)\frac{K(h_n(t)-\xi,t-\tau)}{K(s(t_n)-\xi,t_n-\tau)}d\xi d\tau$$

$$\leq P_n(t)B_n.$$

Note that if $(\xi,\tau)\in G_n$ and $t>t_{nk}$, then $K(h_n(t)-\xi,t-\tau)<e^{-k}$. It follows that there exists a constant M_2, independent of n, such that if $t\in(t_{nk},t_{n,k+1})$, then $P_n(t)\leq e^{-k}M_2$. Hence, $[II_{21}]\leq e^{-rk}M_2B_n$. Finally consider $[II_{22}]$. Let $\mathfrak{M}_n=$measure $H_n\cap(t_{n-1},t_n)$. Note that $\mathfrak{M}_n\to 0$ as $n\to\infty$. This follows because $B_n\to 0$ as $n\to\infty$. Hence,

$$[II_{22}]\leq\int_{H_n\cap(t_{n-1},t_n)}\int_{-\infty}^{\infty}K(h_n(t)-\xi,t-\tau)\,d\xi d\tau$$

$$\leq\int_{H_n\cap(t_{n-1},t_n)}e^{-(t-\tau)}d\tau\leq\int_{t_n-\mathfrak{M}_n}^{t_n}e^{-(t-\tau)}d\tau$$

$$\leq e^{-k}[1-e^{-\mathfrak{M}_n}].$$

Setting $\overline{M}_n=1-\mathfrak{M}_n$, we have that $[II_{22}]\leq\overline{M}_ne^{-rk}$. Note that $\overline{M}_n\to 0$ as $n\to\infty$.

Combining all of these estimates, we have shown that

$$\Phi(h_n)(t)>a-3e^{-rk}e^{-rt_n}+\delta_{nk}M_1-e^{rk}[M_2B_n+\overline{M}_n].$$

Hence $\Phi(h_n)(t)>a$ if we set

$$\delta_{nk}=\frac{e^{-rk}}{M_1}[3e^{-rt_n}+M_2B_n+\overline{M}_n]\equiv K_ne^{-rk}.$$

Note that $K_n\to 0$ as $n\to\infty$. An immediate consequence is that if $\delta_n=\sum_{k=0}^{\infty}\delta_{nk}$, then $\delta_n\to 0$ as $n\to\infty$.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Proc. Tulane Program in Partial Differential Equations and Related Topics, Lecture Notes in Mathematics 446, Springer, Berlin, 1975, pp. 5–49.

[2] _____, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33–76.

[3] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to traveling front solutions*, Arch. Rat. Mech. Anal., 65 (1975), 335–361; Bull. Amer. Math. Soc., 81 (1975), pp. 1075–1078.

[4] R. A. Fisher, *The wave of advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 353–369.

[5] R. FitzHugh, *Impulses and physiological states in models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

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

[7] C. K. R. T. Jones, Ph.D. thesis (1979), Univ. Wisconsin, Madison; MRC Tech. Sum. Rep. 2046, Mathematics Research Center, Univ. Wisconsin, Madison, 1980.

[8] Ya. I. Kanel, *On the stabilization of the Cauchy problem for the equations arising in the theory of combustion*, Mat. Sbornik, 59 (1962), pp. 245–288.

[9] A. N. Kolmogorov, I. G. Petrovskii and N. S. Piskunov, *A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem*, Bjul. Moskovskovo Gos. Univ., 17 (1937), pp. 1–72. (In Russian.)

[10] H. P. McKean, *Nagumo's equation*, Adv. in Math., 4 (1970), pp. 209–223.
[11] J. Nagumo, S. Aromoto and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. Inst. Radio Eng., 50 (1962), pp. 2061–2070.
[12] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[13] D. Terman, *Threshold phenomena in nonlinear diffusion equation*, Ph.D. thesis, Univ. Minnesota, Minneapolis, 1980.
[14] _____, *Local existence for the Cauchy problem of a reaction-diffusion system with discontinuous nonlinearity*, MRC Tech. Sum. Rep. 2221, Mathematics Research Center, Univ. Wisconsin, Madison, 1981.
[15] _____, *Threshold phenomena for a reaction diffusion system*, MRC Tech. Sum. Rep. 2252, Mathematics Research Center, Univ. Wisconsin, Madison, 1981, J. Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0005 $01.25/0

STEADY STATES OF A SYSTEM OF PARTIAL DIFFERENTIAL EQUATIONS MODELING MICROBIAL ECOLOGY*

SZE-BI HSU†

Abstract. In this paper we discuss the existence and uniqueness of solutions for the boundary value problem

$$\begin{aligned} &u''(x)=F(u(x))v(x),\\ &\lambda v''(x)=-[\kappa F(u(x))-\theta]v(x),\\ &u'(0)=0, \quad u(1)=1,\\ &v'(0)=0, \quad v'(1)=0, \end{aligned} \qquad 0\le x\le 1, \quad \lambda,\kappa,\theta>0,$$

which arises in microbial ecology. The growth rate $F(u)$ of bacteria satisfies $F(0)=0$, $F'(u)>0$. We study this problem by using Rabinowitz's global bifurcation theorem and the maximum principle.

1. Introduction. In [1], D. Lauffenburger, R. Aris and K. Keller study the effects of random motility on growth of bacterial populations. Consider a population of bacterial cells confined to a finite region, with a diffusible chemical substrate present in the medium. This substrate is assumed to be the nutrient that is rate limiting for growth, and it is further assumed that it enters the region at a boundary. For simplicity, we consider one-dimensional geometry, with uniform conditions in the transverse dimensions, so that the cells are confined to the region $0\le x\le L$. Substrate enters the region at the boundary $x=L$, and is present there at a constant concentration determined by ambient conditions. We assume Monod's model for the growth of bacterial populations along with exponential nonviability or death. Then the model equations are

$$\begin{aligned} &\frac{\partial b}{\partial t}=\mu\frac{\partial^2 b}{\partial x^2}+[f(s)-k_e]b,\\ &\frac{\partial s}{\partial t}=D\frac{\partial^2 s}{\partial x^2}-\frac{1}{Y}f(x)b \end{aligned} \tag{1.1}$$

for $0\le x\le L$. The boundary conditions are

$$\begin{aligned} &\frac{\partial b}{\partial x}=0, \quad s=s_0 \quad \text{at } x=L,\\ &\frac{\partial b}{\partial x}=0, \quad \frac{\partial s}{\partial x}=0 \quad \text{at } x=0. \end{aligned} \tag{1.2}$$

Here:

$f(s)=ms/(K+s)$,

$b(x,t)=$bacterial cell density at position x and time t (mass of viable cells per volume of medium),

$s(x,t)=$substrate concentration at position x and time t (mole of substrate per volume of medium),

$\mu=$random motility coefficient of bacterial cells,

$D=$substrate diffusion coefficient,

*Received by the editors July 9, 1982. This research was partially supported by the National Science Council of the Republic of China.

†Department of Applied Mathematics, Chiao-Tung University, Hsin-Chu, Taiwan.

k_e = death rate of bacteria population,
Y = yield coefficient (mass of viable cells produced per mole of substrate),
s_0 = constant concentration of substrate present at boundary $x=L$,
m = maximal growth rate of bacterial cells,
K = the half-saturation constant.

Introducing new dimensionless parameters

$$u=\frac{s}{s^{(0)}},\quad \xi=\frac{x}{L},\quad \tau=\frac{Dt}{L^2},\quad v=\frac{bmL^2}{Ys^{(0)}D},\quad \theta=\frac{L^2}{D}k_e,\quad \lambda=\frac{\mu}{D},$$

$$\kappa=\frac{L^2}{D}m,\quad F(u)=\frac{1}{m}f(s^{(0)}u)=\frac{u}{K/s^{(0)}+u}$$

yields equations

$$\frac{\partial u}{\partial \tau}=\frac{\partial^2 u}{\partial \xi^2}-F(u)v,\qquad \frac{\partial v}{\partial \tau}=\lambda\frac{\partial^2 v}{\partial \xi^2}+(\kappa F(u)-\theta)v \tag{1.3}$$

with boundary conditions

$$u(1,\tau)=1,\qquad \frac{\partial u}{\partial \xi}(0,\tau)=0,$$
$$\frac{\partial v}{\partial \xi}(1,\tau)=0,\qquad \frac{\partial v}{\partial \xi}(0,\tau)=0. \tag{1.4}$$

In [1] the authors assume $F(u)=1$ for $u>u_c$ and 0 for $u\le u_c$, where $F(u_c)=\theta/\kappa$, and compute the steady states of (1.3), (1.4). That is, they try to solve the nonlinear problem (1.3), (1.4) by linear techniques. The main purpose of this paper is to show the existence and uniqueness of steady states of (1.3), (1.4). Our technique is to apply the global bifurcation theorem of Rabinowitz [5] and the maximum principle [6].

2. Statements of main results. Consider the steady state problems of (1.3), (1.4)

$$u''(x)=F(u(x))v(x),$$
$$v''(x)=-(\kappa F(u(x))-\theta)v(x), \tag{2.1}$$

for $0\le x\le 1$ with boundary condition

$$u(1)=1,\qquad u'(0)=0,$$
$$v'(1)=0,\qquad v'(0)=0. \tag{2.2}$$

We may assume that $F(u)$ satisfies

$$F(0)=0,\qquad F'(u)>0 \quad \text{for } u>0.$$

Our main result is the following theorem.

THEOREM 2.1. (i) *If* $\kappa F(1)-\theta<0$ *then the trivial solution* $(u_0(x), v_0(x))$ *of* (2.1), (2.2) *is the unique nonnegative solution where* $u_0(x)\equiv 1$, $v_0(x)\equiv 0$.

(ii) *If* $\kappa F(1)-\theta>0$, *then there exists a unique solution* $(u(x), v(x))$ *of* (2.1), (2.2) *with* $u(x)>0$, $v(x)>0$ *for* $0\le x\le 1$.

3. Proof. Our approach is very similar to that of Cushing [2] and Butler et al. [3]. Before we prove our main theorem, we note the following lemmas.

LEMMA 3.1. *Let* $(u(x), v(x))$ *be a solution of* (2.1) *and* (2.2) *with* $u(x)\ge 0$, $v(x)\ge 0$, $0\le x\le 1$. *Then*

(i) $0\le u(x)\le 1$.

(ii) *If* $(u,v)\not\equiv(u_0,v_0)$ *and* $\kappa F(1)-\theta>0$, *then* $u(x)$ *is a strictly convex and strictly increasing function on* $0\le x\le 1$ *while* $v(x)$ *is a strictly increasing function on* $0\le x\le 1$, *and there exists* $0<x_0<1$ *such that* $v(x)$ *is strictly convex on* $(0,x_0)$ *and strictly concave on* $(x_0,1)$.

Proof. From $u'(0)=0$, $u(1)=1$ and $u''\ge 0$ (i) follows easily. If $(u,v)\not\equiv(u_0,v_0)$, then obviously $u(0)\ne 1$; otherwise $u\equiv 1$ and $v\equiv 0$. From the uniqueness of solutions of ODE's and the first equation of (2.1), $u(0)\ne 0$. Hence $u(x)>0$ for $0\le x\le 1$. We claim $v(x)>0$ for $0\le x\le 1$. From the uniqueness of solutions of ODE's and the second equation of (2.1), $v(0)>0$. Suppose the claim is not true. Then there exists $0<\xi<1$ such that $v(\xi)=0$ and $v'(\xi)=0$. Then $v(x)\equiv 0$, and this is the desired contradiction. Hence $u''>0$ on $(0,1)$ and $u(x)$ is a strictly convex and strictly increasing function on $0\le x\le 1$. Obviously it is impossible to have $\kappa F(u(x))-\theta>0$ for all $0\le x\le 1$, since then $v''(x)<0$ for $0\le x\le 1$, which contradicts to the boundary conditions $v'(0)=0=v'(1)$. Hence there exists a unique x_0, $0<x_0<1$, such that $\kappa F(u(x_0))-\theta=0$ and $v''(x)>0$ for $0<x<x_0$, $v''(x)<0$ for $x_0<x\le 1$. Obviously $v(x)$ is strictly increasing on $[0,1]$.

Proof of Theorem 2.1(i). Suppose $(u(x), v(x))$ is a nonnegative steady state, $(u,v)\not\equiv(u_0,v_0)$. Then $u(x)\not\equiv 1$ and $v(x)\not\equiv 0$. From the second equation of (2.1), boundary conditions $v'(0)=v'(1)=0$ and Lemma 3.1 (i), it follows that

$$0=-\int_0^1 v(x)[\kappa F(u(x))-\theta]\,dx>-\int_0^1 v(x)[\kappa F(1)-\theta]\,dx>0.$$

This is a contradiction. Hence we complete the proof.

Before we prove the second part of Theorem 2.1, we need to state the local and global bifurcation theorems, respectively, due to Krasnoselskii [4] and Rabinowitz [5].

LEMMA 3.2 [4]. *Let* $T_\lambda=\lambda A+D$ *be a continuous one-parameter family of operators from a Banach space* X *to itself, such that* A *is compact and linear and satisfies* $\|Dx-Dy\|=o(\|x-y\|)$. *Then a bifurcation of the equation* $T_\lambda x=x$ $(x\in X)$ *can only occur at characteristic value* λ^* (*reciprocal of a nonzero eigenvalue*) *of* A, *and will occur if* λ^* *has odd multiplicity. In this case, the bifurcation point corresponds to a continuous branch of eigenvectors of* T_λ *in a neighborhood of the zero of* X.

LEMMA 3.3 [5]. *Let* T_λ, A, D, X *be as above, and let* S *be the closure of the set of all nontrivial solutions of* $T_\lambda x=x$ *as* λ *ranges over* $\mathbb{R}$. *If* λ^* *is a simple characteristic value of* A, *then* S *contains two subcontinua* C_∞^+, C_∞^- *whose only point in common for* λ *near* λ^* *is* $(\lambda^*,0)$, *and each of which either*

(a) *is unbounded, or*

(b) *contains* $(\hat\lambda,0)$ *where* $\hat\lambda\ne\lambda^*$ *is a characteristic value of* A.

LEMMA 3.4. *For any positive solution* (u,v) *of* (2.1) *we have*

$$v(0)\ge\frac{\kappa}{\lambda}\left[\frac{2((\lambda/\kappa)v(1)+1)}{e^{\sqrt{\alpha}}+e^{-\sqrt{\alpha}}}-u(0)\right]\quad \text{where } \alpha=\frac{\theta}{\lambda}.$$

Proof. From (2.1) we have the following inequality:

$$u''+\frac{\lambda}{\kappa}v''=\frac{\theta}{\kappa}v<\alpha\left(u+\frac{\lambda}{\kappa}v\right), \quad \text{where } \alpha=\frac{\theta}{\lambda}, \tag{3.1}$$

$$\left(u+\frac{\lambda}{\kappa}v\right)'(0)=0,$$

$$\left(u+\frac{\lambda}{\kappa}v\right)(1)=1+\frac{\lambda}{\kappa}v(1).$$

Comparing (3.1) with the equations

$$U''=\alpha U,$$

$$U'(0)=0, \qquad U(1)=1+\frac{\lambda}{\kappa}v(1)$$

yields

$$U(x)=\left(\frac{(\lambda/\kappa)v(1)+1}{e^{\sqrt{\alpha}}+e^{-\sqrt{\alpha}}}\right)(e^{\sqrt{\alpha}x}+e^{-\sqrt{\alpha}x})\leq u(x)+\frac{\lambda}{\kappa}v(x), \qquad 0\leq x\leq 0;$$

in particular,

$$v(0)\geq\frac{\kappa}{\lambda}\left[\frac{2((\kappa/\lambda)v(1)+1)}{e^{\sqrt{\alpha}}+e^{-\sqrt{\alpha}}}-u(0)\right]\geq\frac{\kappa}{\lambda}\left[\frac{2((\lambda/\kappa)v(1)+1)}{e^{\sqrt{\alpha}}+e^{-\sqrt{\alpha}}}-1\right].$$

Proof of Theorem 2.1 (ii) (*existence*). Setting $U=u-u_0$, $V=v-v_0$ in (2.1), we have for $0\leq x\leq 1$,

$$U''=F(1)V+g_1(U,V),$$

$$\lambda V''=\theta V-\kappa F(1)V+g_2(U,V), \tag{3.2}$$

$$U(1)=0, \quad U'(0)=0, \quad V'(0)=0, \quad V'(1)=0$$

where $g_1(U,V)=o(\|(U,V)\|)$, $g_2(U,V)=o(\|(U,V)\|)$ as $(U,V)\to(0,0)$. Consider the linear system

$$U''=F(1)V,$$

$$\lambda V''=\theta V, \qquad 0\leq x\leq 1, \tag{3.3}$$

$$U(1)=0, \quad U'(0)=0, \quad V'(0)=0, \quad V'(1)=0.$$

It is easy to show that (3.3) has only the trivial solution $U\equiv 0$, $V\equiv 0$. Let B be the Banach space of continuous function on $0\leq x\leq 1$ with the supremum norm. If h_1, $h_2\in B$, let $L_1(h_1)$, $L_2(h_2)$ respectively, be the unique solutions of

$$U''=h_1, \quad U'(0)=0, \qquad U(1)=0, \tag{3.4}$$

$$\lambda V''=\theta V+h_2, \quad V'(0)=0, \quad V'(1)=0. \tag{3.5}$$

Obviously L_1, $L_2: B\to B$ are linear and compact operators.

Write (3.2) formally as the following operator equation:

$$\begin{pmatrix} U \\ V \end{pmatrix}=\kappa L^*\begin{pmatrix} U \\ V \end{pmatrix}+G\begin{pmatrix} U \\ V \end{pmatrix}, \tag{3.6}$$

where

$$L^*\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} -F(1)L_1 \circ L_2(F(1)V) \\ L_2(-F(1)V) \end{pmatrix},$$

$$G\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} F(1)L_1 \circ L_2(g_2(U,V)) + L_1(g_1(U,V)) \\ L_2(g_2(U,V)) \end{pmatrix}$$

and $L^*: B\times B \to B\times B$ is compact and linear while $G: B\times B \to B\times B$ is compact and $G(U,V) = o(\|(U,V)\|)$ as $\|(U,V)\| \to 0$. We now formally treat κ in (3.5) as a real parameter. Consider the eigenvalue problem

$$\begin{pmatrix} U \\ V \end{pmatrix} = \kappa L^*\begin{pmatrix} U \\ V \end{pmatrix}. \tag{3.7}$$

CLAIM. *The characteristic values of* L^* *are* $\kappa^* = \theta/F(1)$ *and*

$$\kappa_n = \frac{\theta + \lambda(n\pi)^2}{F(1)}, \qquad n = 1, 2, \cdots.$$

Let κ be a characteristic value of L^*. Then there exists $\binom{U}{V} \neq \binom{0}{0}$ such that

$$\begin{pmatrix} U \\ V \end{pmatrix} = \kappa L^*\begin{pmatrix} U \\ V \end{pmatrix} = \kappa\begin{pmatrix} -F(1)L_1 \circ L_2(F(1)V) \\ L_2(-F(1)V) \end{pmatrix}$$

or the system

$$\begin{aligned} &U'' = F(1)V, \\ &\lambda V'' = \theta V - \kappa F(1)V, \\ &U(1) = 0, \quad U'(0) = 0, \quad V'(0) = 0, \quad V'(1) = 0 \end{aligned} \tag{3.8}$$

has nontrivial solutions.

If $\kappa F(1) - \theta < 0$ then $U \equiv 0$, $V \equiv 0$.

If $\kappa F(1) - \theta = 0$ then the eigenspace belonging to $(\kappa^*)^{-1} = (\theta/F(1))^{-1}$ is generated by (U_1, V_1), where $U_1(x) = (F(1)/2)(x^2 - 1)$ and $V_1(x) \equiv 1$. If $\kappa F(1) - \theta > 0$ then $V'' + \alpha V = 0$, $\alpha = (\kappa F(1) - \theta)/\lambda > 0$, $V'(0) = 0 = V'(1)$.

In order to have $V \not\equiv 0$, α must satisfy $\sqrt{\alpha} = n\pi$ and $V(x) = C\cos n\pi x$, $U(x) = (-F(1)C/(n\pi)^2)\cos n\pi x$ where C is an arbitrary constant. Hence the eigenspace belonging to $(\kappa_n)^{-1} = ((\theta + \lambda(n\pi)^2)/F(1))^{-1}$ is generated by (U_n, V_n), $U_n(x) = -F(1)/(n\pi)^2 \cdot \cos n\pi x$, $V_n(x) = \cos n\pi x$.

By Lemma 3.2, bifurcation does indeed occur for $\kappa = \kappa^*$, and we obtain a continuous branch of solutions of (3.6) all of which are nontrivial except for the solution $(\kappa^*, 0, 0)$. A Lyapunov–Schmidt series expansion of these solution (κ, U, V) near $(\kappa^*, 0, 0)$ reveals that we have solutions of (3.6) that correspond to the positive solutions of (2.1). In fact, let

$$\begin{aligned} U(x) &= \varepsilon\tilde{U}_1(x) + \varepsilon^2\tilde{U}_2(x) + \varepsilon^3\tilde{U}_3(x) + \cdots, \\ V(x) &= \varepsilon\tilde{V}_1(x) + \varepsilon^2\tilde{v}_2(x) + \varepsilon^3\tilde{v}_3(x) + \cdots, \\ \kappa &= \kappa^* + \tilde{\kappa}_1\varepsilon + \tilde{\kappa}_2\varepsilon^2 + \cdots \end{aligned} \tag{3.9}$$

and we find that

$$\tilde{U}_1''(x)=F(1)\tilde{V}_1(x),$$
$$\tilde{V}_1''(x)=\tilde{V}_1(x)-\kappa^*F(1)\tilde{V}_1(x),$$
$$\tilde{U}_1(1)=0,\quad \tilde{U}_1'(0)=0,\quad \tilde{V}_1'(0)=0,\quad \tilde{V}'(1)=0.$$

Choose $\tilde{U}_1(x)=U_1(x)$, $\tilde{V}_1(x)=V_1(x)$ and obviously $\tilde{\kappa}_1>0$ by Theorem 2.1(i).

To complete the proof for the existence part, we need to show that such a solution exists for all $\kappa>\kappa^*$. Since κ^* is a simple characteristic value of L^*, it follows from Lemma 3.3 that there is a continuum C_∞^+ of solutions of (3.6) all of which are nontrivial except for the solution $(\kappa^*,0,0)$ such that C_∞^+ either is unbounded or contains $(\kappa_n,0,0)$ for some n.

Our approach is first to eliminate the latter possibility. Let D_+^∞ be the nontrivial solutions of (2.1) corresponding to C_+^∞. We claim

$$(\kappa,u,v)\in D_\infty^+ \Rightarrow u>0,\ v>0 \text{ and } \kappa\geq\kappa^*. \tag{3.10}$$

Since $u>0$, $v>0$ near the bifurcation point $(\kappa^*,1,0)$ and C_∞^+ is a continuum. If (3.10) does not hold then by Lemma 3.1 there exists $(\kappa_0,u,v)\in D_\infty^+$ such that $u(0)=0$ or $v(0)=0$. If $u(0)=0$ then from $u''=F(u)v$, $u'(0)=0$ it follows that $u\equiv 0$, which contradicts $u(1)=1$. On the other hand, if $v(0)=0$ then from $v''=-v(\kappa_0F(u)-\theta)$, $v'(0)=0$ it follows that $v\equiv 0$ and hence $u\equiv 1$, which contradicts the fact D_+^∞ does not contain a trivial solution. It is obvious from Lemma 3.1 that $\kappa\geq\kappa^*$.

Now we suppose C_∞^+ contains $(\kappa_n,0,0)$ for some n. A Lyapunov–Schmidt expansion about $(\kappa_n,0,0)$ as in (3.9) reveals

$$U(x)=\varepsilon U_n(x)+\varepsilon^2\tilde{U}_{n,2}(x)+\cdots,$$
$$V(x)=\varepsilon V_n(x)+\varepsilon^2\tilde{V}_{n,2}(x)+\cdots,$$
$$\kappa=\kappa_n+\tilde{\kappa}_{n-1}\varepsilon+\cdots,$$

where $U_n(x)=(-F(1)/(n\pi)^2)\cos n\pi x$, $V_n(x)=\cos n\pi x$, $\kappa_n=(\theta+\lambda(n\pi)^2)/F(1)$. It obviously contradicts (3.10) in a neighborhood of $(\kappa_n,0,0)$. Hence C_∞^+ must be unbounded.

Now let Λ, Y be the projections of D_+^∞ onto the real axis and $B\times B$ respectively. To complete the proof of the existence part we show that

$$\Lambda=[\kappa^*,\infty). \tag{3.11}$$

Suppose (3.11) does not hold. Then we may assume $\Lambda=[\kappa^*,\bar{\kappa}]$ and Y is unbounded. Then there exists a sequence of points $\{(\bar{\kappa}_n,u_n,v_n)\}_{n=1}^\infty$ in D_+^∞ such that $\bar{\kappa}_n\to\bar{\kappa}_0\in\Lambda$ and $\|(u_n,v_n)\|\to\infty$ as $n\to\infty$. Since $|u_n(x)|\leq 1$ for all $0\leq x\leq 1$, by Lemma 3.1 and 3.4 it follows that $v_n(1)\to+\infty$ and $v_n(0)\to+\infty$ as $n\to\infty$. From Lemma 3.4, there exist $N_0\geq 0$, $C>0$ (C is independent of n) such that $v_n(0)>Cv_n(1)$ for all $n\geq N_0$. Now we choose $\varepsilon>0$ sufficiently small that $\bar{\kappa}F(\varepsilon)-\theta<0$ and let

$$x_0=\frac{C(\bar{\kappa}F(1)-\theta)}{C(\bar{\kappa}F(1)-\theta)-(\bar{\kappa}F(\varepsilon)-\theta)};$$

then $0<x_0<1$. We claim:

$$\text{There exists } n>N_0 \text{ such that } u_n(x_0)<\varepsilon. \tag{3.12}$$

If (3.12) does not hold, then $u_n(x_0)>\varepsilon$ for all $n>N_0$ and hence $u_n(x)\geq\varepsilon$ for all $x_0\leq x\leq 1$, $n\geq N_0$. Then $u_n''(x)\geq F(u_n(x))v_n(x)\geq F(\varepsilon)v_n(x)\geq F(\varepsilon)v_n(0)$ for all $x_0\leq x\leq 1$ and $\min_{x_0\leq x\leq 1} u_n''(x)\to+\infty$ as $n\to\infty$. But

$$u_n(1)-u_n(x_0)=u_n'(x_0)\cdot(1-x_0)+\frac{u_n''(\xi)}{2}(1-x_0)^2$$
$$>\Big\{\min_{x_0\leq x\leq 1} u_n''(x)\Big\}\cdot(1-x_0^2)\to+\infty \quad \text{as } n\to\infty,$$

and this contradicts the fact that $0<u_n(1)-u_n(x_0)<1$. Hence we establish (3.12).

Consider n as in (3.12). By the second equation in (2.1) we have

$$\int_0^1 v_n(x)\big[\kappa_n F(u_n(x))-\theta\big]\,dx=0. \tag{3.13}$$

Let

$$\text{L.H.S. of (3.13)}=\int_0^{x_0} v_n(x)\big[\kappa_n F(u_n(x))-\theta\big]\,dx+\int_{x_0}^1 v_u(x)\big[\kappa_n F(u_n(x))-\theta\big]\,dx.$$

Then

$$\begin{aligned}
0&<\int_0^{x_0} v_n(x)\big[\bar\kappa F(\varepsilon)-\theta\big]\,dx+\int_{x_0}^1 v_n(x)\big[\kappa_n F(1)-\theta\big]\,dx\\
&<v_n(0)(\bar\kappa F(\varepsilon)-\theta)x_0+(1-x_0)v_n(1)(\bar\kappa F(1)-\theta)\\
&<v_n(0)(\bar\kappa F(\varepsilon)-\theta)x_0+C(1-x_0)(\bar\kappa F(1)-\theta)v_n(0)\\
&=v_n(0)\big[C(\bar\kappa F(1)-\theta)-x_0(C(\bar\kappa F(1)-\theta)-(\bar\kappa F(\varepsilon)-\theta))\big]\\
&=0.
\end{aligned}$$

Hence we obtain the desired contradiction and (3.11) holds. Q.E.D.

Our next step is to show the uniqueness of the nonnegative solution of (2.1), (2.2). Before we prove it, we present the following lemmas.

LEMMA 3.5. *Let* (u_1,v_1), (u_2,v_2) *be nonnegative solutions of* (2.1) *and* (2.2) *with* $u_1\geq u_2$. *Then* $u_1\equiv u_2$, $v_1\equiv v_2$.

Proof. Suppose $u_1\geq u_2$ and $u_1\not\equiv u_2$. Let $\omega=v_2/v_1$. Then from (2.1), (2.2) we have

$$\omega''+2\left(\frac{v_1'}{v_1}\right)\omega'+\omega\big[\kappa(F(u_2)-F(u_1))\big]=0, \tag{3.14}$$
$$\omega'(0)=0, \qquad \omega'(1)=0.$$

Since $F(u_2)-F(u_1)\leq 0$, from the maximum principle [6] it follows that $\omega\equiv\text{constant}>0$. But from (3.14) and $u_1\not\equiv u_2$, we have a contradiction. Hence $u_1\equiv u_2$ and $v_1\equiv v_2$.

LEMMA 3.6. *Let* (u_1,v_1), (u_2,v_2) *be nonnegative solutions of* (2.1), (2.2) *with* $u_1\not\equiv u_2$. *Then the curve* $y=u_1(x)$ *crosses the curve* $y=u_2(x)$ *a finite number of times on* $0\leq x\leq 1$.

Proof. From Lemma 3.5, the curve $y=u_1(x)$ must cross the curve $y=u_2(x)$ on $0\leq x\leq 1$. Suppose $y=u_1(x)$ crosses the curve $y=u_2(x)$ an infinite number of times on $0\leq x\leq 1$. Then there exists $\{x_n\}_{n=1}^{\infty}$ such that $u_1(x_n)=u_2(x_n)$ and there exists $a\in[0,1]$ such that $x_n\to a$ as $n\to\infty$. Obviously $u_1(a)=u_2(a)$. Let $U(x)=u_1(x)-u_2(x)$, $0\leq x\leq 1$. Since for any neighborhood of a, the curve $y=u_1(x)$ crosses $y=u_2(x)$ an infinite

number of times, the Taylor expansion of $U(x)$ at a yields $U'(a)=0$, $U''(a)=0$, $U'''(a)=0$. Hence $u_1'''(a)=u_2'(a)$, $u_1''(a)=u_2''(a)$, $u_1'''(a)=u_2''(a)$. From (2.1) we have $v_1(a)=v_2(a)$, $v_1'(a)=v_2'(a)$, $v_1''(a)=v_2''(a)$. However the uniqueness of the solution of the ordinary differential equations (2.1) yields $u_1\equiv u_2$, $v_1\equiv v_2$. Hence we complete the proof of the lemma.

Proof of Theorem 2.1(ii) (*uniqueness*). Suppose we have two nonnegative solutions of (2.1), (2.1), say (u_1,v_1), (u_2,v_2) with $u_1\not\equiv u_2$ under the assumption $\kappa F(1)-\theta>0$. By Lemma 3.1(ii) u_1, u_2, v_1, v_2 are positive on $0\le x\le 1$. From Lemmas 3.5, 3.6, the curve $y=u_1(x)$ crosses the curve $y=u_2(x)$ a finite number of times. Let $x_0=0$, $x_{n+1}=1$ and $x_1,\cdots,x_n$ be the points where two curves cross each other. Without loss of generality, we may assume $u_1\ge u_2$ on $[x_k,x_{k+1}]$, where $0\le k\le n$, k even, and $u_2\ge u_1$ on $[x_k,x_{k+1}]$ where $0\le k\le n$, k odd. In order to obtain a contradiction, we discuss two cases.

Case 1. $v_1(0)\le v_2(0)$. Let $\omega=v_2/v_1$ on $0\le x\le x_1$. Then we have

$$\omega''+2\left(\frac{v_1'}{v_1}\right)\omega'+\omega\left[\kappa\left(F(u_2)-F(u_1)\right)\right]=0,\qquad \omega'(0)=0. \tag{3.15}$$

Then the maximum principle yields $v_2(x)>v_1(x)$ for all $0<x\le x_1$. We claim $y=v_2(x)$ must cross $y=v_1(x)$ at some point $c_1\in(x_1,x_2)$. If not, then $v_2\ge v_1$, $u_2\ge u_1$ on $[x_1,x_2]$. Since $u_2(x_2)=u_1(x_2)$ and $u_2'(x_1)\ge u_1'(x_1)$, $u_2(x_1)=u_1(x_1)$, it follows that

$$\begin{aligned} u_1(x_2)=u_2(x_2)&=u_2(x_1)+(x_2-x_1)u_2'(x_1)+\int_{x_1}^{x_2}\int_{x_1}^{s}F(u_2(\eta))v_2(\eta)\,d\eta\,ds\\ &>u_1(x_1)+(x_2-x_1)u_1'(x_1)+\int_{x_1}^{x_2}\int_{x_1}^{s}F(u_1(\eta))v_1(\eta)\,d\eta\,ds\\ &=u_1(x_2). \end{aligned}$$

This is a contradiction. Similarly, let $\bar\omega=v_1/v_2$ on $c_1\le x\le x_2$. Then

$$\bar\omega''+2\left(\frac{v_2'}{v_2}\right)\bar\omega'+\bar\omega\left[\kappa\left(F(u_1)-F(u_2)\right)\right]=0,\qquad \bar\omega(c_1)=1. \tag{3.16}$$

The maximum principle yields $v_1>v_2$ on $(c_1,x_2]$.

Repeating the arguments shows that there exist $c_2,\cdots,c_n$, $x_i<c_i<x_{i+1}$, $i=1,\cdots,n$ such that $v_1(c_i)=v_2(c_i)$, $i=1,\cdots,n$, $v_1\ge v_2$ on $[c_i,c_{i+1}]$ where i is odd, and $v_2\ge v_1$ on $[c_i,c_{i+1}]$ where i is even. If $u_1\ge u_2$ on $[x_n,1]$ then $v_2\ge v_1$ on $[c_n,1]$. Consider (3.15) on $[c_n,1]$; then the maximum of $\omega=v_2/v_1$ occurs at $x=1$ but $\omega'(1)=0$ and we obtain a contradiction. If $u_2\ge u_1$ on $[x_n,1]$ then $v_1\ge v_2$ on $[c_n,1]$. Similarly, consider (3.16) on $[c_n,1]$; the maximum of $\bar\omega=v_1/v_2$ occurs at $x=1$, but $\bar\omega'(1)=0$ and we obtain a contradiction.

Case 2. $v_2(0)<v_1(0)$. We claim that the curve $y=v_1(x)$ must cross $y=v_2(x)$ at some point $\bar c_0\in(0,x_1)$. If not, then $u_1\ge u_2$, $v_1\ge v_2$ on $[0,x_1]$. Since $u_1(0)\ge u_2(0)$, $u_1'(0)=u_2'(0)=0$, we have

$$\begin{aligned} u_2(x_1)=u_1(x_1)&=u_1(0)+\int_0^{x_1}\int_0^{s}F(u_1(\eta))v_1(\eta)\,d\eta\,ds\\ &>u_2(0)+\int_0^{x_1}\int_0^{s}F(u_2(\eta))v_2(\eta)\,d\eta\,ds=u_2(x_1). \end{aligned}$$

By the arguments in Case 1, there exist $\bar{c}_1, \cdots, \bar{c}_n$ such that $x_i < \bar{c}_i < x_{i+1}$, $i = 1, \cdots, n$ such that $v_1(\bar{c}_i) = v_2(\bar{c}_i)$ and v_1, v_2 cross each other at $\bar{c}_i$. Applying the same arguments as in Case 1 we obtain a contradiction.

Hence we establish the uniqueness of solutions for (2.1), (2.2).

Discussion. We have established the existence and uniqueness of steady states for the equations (1.3), (1.4). As for the questions about the global behavior of solutions for this dynamical system, it is currently under investigation. From our numerical studies, the steady state should be globally asymptotically stable. This paper is the first step in discussing the effects of motility in the model studied in [1] which will provide a reasonable explanation for the phenomena in microbial ecology.

REFERENCES

[1] D. LAUFFENBURGER, R. ARIS AND K. KELLER, *Effects of random motility on growth of bacterial populations*, Microb. Ecol., (1981), pp. 207–227.

[2] J. M. CUSHING, *Periodic time-dependent predator-prey systems*, SIAM J. Appl. Math., 32 (1977), pp. 82–95.

[3] G. J. BUTLER AND H. I. FRELDMAN, *Periodic solutios of a predator-prey system with periodic coefficients*, Math. Biosc., 55 (1981), pp. 27–38.

[4] M. A. KRASNOSELSKII, *Topological Methods in the Theory of Nonlinear Integral Equations*. Macmillan, New York, 1964.

[5] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Functional Analysis, 7 (1971), pp. 487–513.

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

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0006 $01.25/0

THE QUENCHING OF SOLUTIONS OF LINEAR PARABOLIC AND HYPERBOLIC EQUATIONS WITH NONLINEAR BOUNDARY CONDITIONS*

HOWARD A. LEVINE†

Abstract. In this paper we examine the initial-boundary value problems (α): $u_t=u_{xx}$, $0<x<L$, $t>0$, $u(0,t)=u(x,0)=0$, $u_x(L,t)=\phi(u(L,t))$ and (β): $u_{tt}=u_{xx}$, $0<x<L$, $t>0$, $u(0,t)=u(x,0)=u_t(x,0)$, $u_x(L,t)=\phi(u(L,t))$ where $\phi(-\infty,1)\to(0,\infty)$ is continuously differentiable, monotone increasing and $\lim_{u\to 1^-}\phi(u)=+\infty$. For problem ($\alpha$) we show that there is a positive number L_0 such that if $L\le L_0$, $u(x,t)\le 1-\delta$ for some $\delta>0$ for all $t>0$, while if $L>L_0$, $u(L,t)$ reaches one in finite time while $u_t(L,t)$ becomes unbounded in that time. For problem (β) it is shown that if L is sufficiently small, then $u(L,t)\le 1-\delta$ for all $t>0$ while if L is sufficiently large and $\int_0^1\phi(\eta)\,d\eta<\infty$, $u(L,t)$ reaches one in finite time whereas if $\int_0^1\phi(\eta)\,d\eta=\infty$, $u(L,t)$ reaches one in finite or infinite time.

In either of the last two situations $u_t(L,t)$ becomes unbounded if the time interval is finite. If u reaches one in infinite time, then $\int_0^1 u_x^2(x,t)\,dx$ and $u(x,t)$ are unbounded on the half line and half strip respectively.

1. Introduction.

In his paper [5], Kawarada studied the behavior of solutions of

$$\text{(A)}\qquad \begin{aligned} &u_t=u_{xx}+1/(1-u(x,t)), && t>0,\quad 0<x<L,\\ &u(0,t)=u(L,t)=0, && t>0,\\ &u(x,0)=0, && 0\le x\le L. \end{aligned}$$

He showed that if $u(L/2,t)$ reached one in finite time, T, then $u_t(L/2,t)$ was unbounded on $(0,T)$, in fact $\lim_{t\to T^-}u_t(L/2,t)=+\infty$. He called this type of regularity loss quenching. In the same paper, he showed that if $L>2\sqrt{2}$, then quenching must occur as $u(L/2,t)$ does then reach one in finite time. In [1,2] and independently in [6] it was shown that there is a number $L_0<2\sqrt{2}$ ($L_0\cong 1.5307$) such that if $L<L_0$ then u cannot quench, even in infinite time whereas if $L>L_0$ u must quench in finite time. In [6] it was also shown that if $L=L_0$ the former situation holds. In [1], [2], [6] more general nonlinearities were also studied.

Let us make the following operational definition, which is weaker than Kawarada's. We will say that a solution of an evolutionary equation quenches in some seminorm (in x) depending on t if (i) the solution remains bounded in this norm while (ii) some derivative in some seminorm of the solution becomes unbounded *in finite time*. We shall sometimes say that a solution quenches in infinite time if (i) and (ii) occur but the solution exists on $[0,L]\times[0,\infty)$.

In [3] the following nonlinear initial boundary value problem for the wave equation was studied.

$$\text{(B)}\qquad \begin{aligned} &u_{tt}=u_{xx}+1/(1-u), && t>0,\quad 0<x<L,\\ &u(0,t)=u(L,t)=0, && t>0,\\ &u(x,0)=u_t(x,0)=0, && 0<x<L. \end{aligned}$$

The interest in (B) was theoretical. Whereas for (A) heavy use of the maximum principle was made, for (B) it was necessary to employ other arguments, specifically

*Received by the editors October 12, 1981, and in revised form July 12, 1982. This research was supported in part by the Science and Humanities Research Institute of the Iowa State University of Science and Technology.

†Department of Mathematics, Iowa State University, Ames, Iowa 50011.

energy arguments, to establish global existence and no quenching, even in infinite time, for small L and a differential inequality argument to establish quenching for large L.

Although our interest in (A) and (B) was theoretical, both problems have their origin in physics. Problem (A) arises in the study of electric current transients in polarized ionic conductors [5]. Problem (B) can be viewed as the initial-boundary value problem describing the motion of a wire composed of a magnetic material carrying an electric current in the presence of a second wire also carrying a current. Stoker and Minorsky [9], [10] give a phase plane analysis of the analogous ordinary differential equation which describes the motion of a current carrying conductor restrained by springs and subject to the force due to a magnetic field of an infinitely long parallel wire conducting a current I. The equation has the form

$$\ddot{x}(t) = -kx(t) + k\lambda/(a - x(t))$$

where $x(0)$, $\dot{x}(0)$ are prescribed.

We were aware of the physical motivation for (A) before we wrote [6]. However Arje Nachman kindly brought to our attention the references [9], [10] (unfortunately after [3] had appeared). In the same spirit and in the hope that the knowledgeable reader will have a ready application for them, we present our results for problems (α), (β) below.

It is the purpose of this paper to examine the corresponding problems when the solution is driven by the boundary conditions rather than by the forcing term. Specifically, we study

(α)
$$\begin{aligned} &u_t = u_{xx}, && t>0, \quad 0<x<L, \\ &u(0,t) = 0, && t \geq 0, \\ &u_x(L,t) = \phi(u(L,t)), && t>0, \\ &u(x,0) = 0, && 0<x \leq L \end{aligned}$$

and

(β)
$$\begin{aligned} &u_{tt} = u_{xx}, && t>0, \quad 0<x<L, \\ &u(0,t) = 0, && t \geq 0, \\ &u_x(L,t) = \phi(u(L,t)), && t>0, \\ &u(x,0) = u_t(x,0) = 0, && 0 \leq x \leq L. \end{aligned}$$

(By simultaneous scaling in x,t we can take $L=1$ in Problems (α), (β) provided the boundary condition at the right endpoint takes the form

$$u_x(1,t) = L\phi(u(1,t)), \qquad t>0.$$

We shall therefore take $L=1$ and use the boundary condition above without further mention of this reduction.) Here ϕ: $(-\infty, 1) \to (0, \infty)$ is continuously differentiable, monotone increasing and $\lim_{u \to 1^-} \phi(u) = +\infty$. The boundary and initial data are taken to be zero, not only for convenience, but also so that one can isolate the effects of the nonlinearity on the solution. While the results for (α), (β) are similar to those obtained for (A), (B), there are several differences worthy of mention. In the first place we show here that not only does, for large L, $u(1,t)$ (problem (α)) become one in finite time but also $u_t(1,t)$ becomes infinite in finite time. The same is true for problem (β) if $\int_0^1 \phi(\eta)\,d\eta < \infty$. If, for ($\beta$), $\int_0^1 \phi(\eta)\,d\eta = +\infty$, then u must quench in finite or infinite time. If this time is finite, $u_t(1,t)$ becomes unbounded in finite time also. If this time is

infinite then u is unbounded on $[0,1]\times[0,\infty)$ and $\int_0^1 u_x^2(x,t)\,dx$ is also unbounded on $[0,\infty)$. This is a somewhat weaker result when $\phi(u)=1/(1-u)$ than for (B). On the other hand, it is shown for both Problems (α), (β) that if L is small then quenching cannot occur for either problem, even in infinite time. The results for problem (α) are sharp while there is a gap for problem (β). That is, for problem (α), there is $L_0>0$ such that if $L\le L_0$ no quenching at all is possible while if $L>L_0$ u quenches. For problem (β) there are L_1, L_2 with $0<L_1<L_2$ such that if $L<L_1$ no quenching at all can occur whereas if $L>L_2$ some kind of quenching must occur. These results are in accord with the general principle that small domains are more stable than large domains.

It is perhaps worth mentioning that, via the change of variable

$$v(x,t)=\int_1^{u(x,t)} d\eta/\phi(\eta)\equiv\Psi(u(x,t)),$$

we may reduce (α) to

$$(\alpha')\qquad \begin{aligned} &v_t=v_{xx}+\phi'(\Psi^{-1}(v))v_x^2, && 0<x<1,\quad t>0,\\ &v(x,0)=\Psi(0), && 0\le x\le 1,\\ &v(0,t)=\Psi(0), && t\ge 0,\\ &v_x(1,t)=L, && t>0. \end{aligned}$$

Using the techniques of [1], [2], [6], it is possible to study the problem

$$(\alpha'')\qquad \begin{aligned} &u_t=u_{xx}+L^2\phi(u), && 0<x<1,\quad t>0,\\ &u(x,0)=0, && 0<x<1,\\ &u(0,t)=0, && t\ge 0,\\ &u_x(1,t)=0, && t\ge 0. \end{aligned}$$

(In [1],[2],[6] the condition at $x=1$ was $u(1,t)=0$.) The same substitution reduces this problem to

$$(\alpha''')\qquad \begin{aligned} &v_t=v_{xx}+\phi'(\Psi^{-1}(v))v_x^2+L^2, && 0<x<1,\quad t>0,\\ &v(x,0)=\Psi(0), && 0<x<1,\\ &v(0,t)=\Psi(0),\\ &v_x(0,t)=0. \end{aligned}$$

Certainly (α') and (α''') are similar looking problems and we might therefore expect (indeed it is our goal to show) that the results obtained for (α) are similar to those obtained for (α''). However, there is no obvious correspondence between the solutions of (α) and (α'') or between (α') and (α'''). For example, the stationary solutions of (α) (when they exist) are linear in x so that (for $\phi(u)=1/(1-u)$) the stationary solutions of (α') are quadratic polynomials in x since

$$\int_1^u (1-\eta)\,d\eta=-\frac{1}{2}(1-u)^2.$$

On the other hand, (again for $\phi(u)=1/(1-u)$) the stationary solutions of (α''') (when they exist) are transcendental functions of x (see [1],[2],[6]). Therefore a separate treatment is needed for (α). Similar remarks apply to problem (β).

The plan of the paper is as follows. In §2, we treat problem (α) first establishing global existence for small L and then quenching for large L. In §3, problem (β) is analyzed in the same manner. In §4 we discuss local existence. The local existence

result for (β) is of special interest. This ordering of topics introduces a slight nonlinearity in the development of our results. We apologize to the reader for this. However, the more interesting results, namely Theorem 2.5, Corollary 2.7, Theorem 3.1 and Corollary 3.4, come first and provide the main thrust of the paper.

A word about notation. We let $D_T=(0,1)\times(0,T)$ if $T<\infty$ and $D=(0,1)\times(0,\infty)$. Likewise, if $T<\infty$, $\Gamma_T=(0,1)\times\{0\}\cup\{0,1\}\times[0,T)$ and $\Gamma=(0,1)\times\{0\}\cup\{0,1\}\times[0,\infty)$ denote the parabolic boundary of D_T and D respectively.

The results of this paper, as well as those of [1]–[6], have some higher dimensional analogues. For example, in [1] and [2], problem (A) was studied in several dimensions. However, the blowup of u_t has yet to be shown for such problems. Likewise, Lieberman and the author have obtained some extensions of the results for problem (α) in several variables, but again with less sharp results than in one dimension. (However, we have an example of infinite time quenching in two dimensions, which cannot occur in one dimension.)

The hyperbolic problems (B), (β) present a more difficult challenge in several space dimensions. For both problems, it is fairly easy to obtain quenching if the space domain is large enough. However, the question of global existence for small domains is open when d^2/dx^2 is replaced by a second order elliptic operator. Smiley and the author have extended the global existence result when d^2/dx^2 is replaced by an elliptic operator of sufficiently high order. The essential ingredient in the global existence argument is the existence of a continuous imbedding of $W_0^{1,p}(\Omega)$ into $L^\infty(\Omega)$ for sufficiently large p, i.e., an inequality of the form

$$|u(x,t)|^2\leq\text{const.}\times\left(\int_\Omega|u(y,t)|^{2p}dy\right)^{1/p}$$

for $\Omega\subset R^n$, Ω bounded, $\partial\Omega$ smooth and $u=0$ on a portion of $\partial\Omega$ of dimension $n-1$. Finally, nothing has been established about the behavior of u_{tt}, even in one dimension.

The above paragraph corrects one of the concluding remarks of [3]. We note one other correction (typographical) for [3]. On p. 395, we should have

$$F(x,t,u)=-\varphi(-u)\quad\text{if } x\in[2n-1,2n).$$

2. The parabolic problem (α). By a solution of (α) in D_T we mean a function $u(x,t)$ continuous in $D_T\cup\Gamma_T$, $u<1$ on $D_T\cup\Gamma_T$, and twice continuously differentiable in x and once in t in D_T. Known regularity results permit differentiation of the equation in D_T. The following lemmas are easy consequences of the maximum principle and the boundary point lemma for parabolic equations (These are sometimes referred to as the first and second maximum principles for parabolic inequalities. See [7, pp. 164, ff.]).

LEMMA 2.1. *If u solves (α) in D_T, then $u>0$ there.*

Proof. For any $\varepsilon>0$, if u had a nonpositive minimum in $\bar{D}_{T-\varepsilon}$, by the maximum principle it would have to occur on $\bar{\Gamma}_{T-\varepsilon}$. Since $u_x(1,t)\geq0$ if $0<t\leq T-\varepsilon$ it cannot occur on line $x=1$. Since u is not identically zero, $u>0$ in $\bar{D}_{T-\varepsilon}$ for all $\varepsilon>0$.

LEMMA 2.2. *If u solves (α) in D_T, then $u_x(x,t)>0$ in D_T.*

Proof. Put $\pi=u_x(x,t)$. Then $\pi_{xx}=\pi_t$ in D_T, $\pi(0,t)\geq0$ by Lemma 2.1, $\pi(x,0)=0$ and $\pi(1,t)>0$ so $\pi>0$ in D_T again by the maximum principle and the boundary point lemma.

LEMMA 2.3. *If u solves (α) in D_T, then $u_t(x,t)>0$ in $D_T\cup\{1\}\times(0,T)$.*

Proof. We work in $D_{T-\varepsilon}$ with $0<h\le\varepsilon/2$ and $0\le t\le T-\varepsilon$. Define $v(x,t)=u(x,t+h)-u(x,t)$. Then v solves

$$\begin{aligned}
&v_t=v_{xx}, && 0<x<1,\quad 0\le t<T-\varepsilon,\\
&v(x,0)>0, && 0<x<1,\\
&v(0,t)=0, && 0\le t\le T-\varepsilon,\\
&v_x(1,t)=L[\phi(u(1,t+h))-\phi(u(1,t))], && 0<t\le T-\varepsilon\\
&\qquad =L\phi'(\xi)v,
\end{aligned}$$

where ξ is between $u(1,t)$ and $u(1,t+h)$. Since $t+h\le T-\varepsilon/2$, $u(1,t)$ and $u(1,t+h)$ are bounded above by $1-\delta'$ for some $\delta'>0$. Therefore there is a number $\lambda<0$ such that $\lambda+L\phi'(\xi)<0$ if $t\le T-\varepsilon$ and $h\le\varepsilon/2$. Now set $w=v\exp(\lambda x-\lambda^2 t)$. Then

$$\begin{aligned}
&w_t-w_{xx}+2\lambda w_x=0 && \text{in } D_{T-\varepsilon},\\
&w(0,t)=0, && 0\le t\le T-\varepsilon,\\
&w(x,0)>0 \text{ (by Lemma 2.1)}, && 0\le x\le L,\\
&w_x(1,t)=(\lambda+L\phi(\xi))w, && 0<t\le T-\varepsilon.
\end{aligned}$$

By the maximum principle, w cannot have a nonpositive minimum in $D_{T-\varepsilon}\cap(0,1)\times\{T-\varepsilon\}$. Moreover w cannot have a negative minimum at a point $(1,t_0)$ $(0<t_0<T-\varepsilon)$, otherwise (because of the choice of λ) $w_x(1,t_0)>0$ at such a point whereas it must be nonpositive at a negative minimum. Therefore $w\ge 0$ in $\bar{D}_{T-\varepsilon}$ and hence $v\ge 0$ in D.

It follows that, wherever it exists, $u_t(x,t)\ge 0$. Now $v=u_t$ satisfies

$$\begin{aligned}
&v_t=v_{xx}, && 0<x<1,\quad 0<t<T-\varepsilon,\\
&v(x,0)\ge 0, && 0\le t\le T-\varepsilon,\\
&v_x(1,t)=L\phi'(u(1,t))v, && 0<T\le T-\varepsilon.
\end{aligned}$$

Since $v\ge 0$, v cannot vanish at any point in the set $\{(x,t)|0\le x<1, 0<t\le T-\varepsilon\}$ unless $v\equiv 0$ by the strong maximum principle. However if $v\equiv 0$, then $u(x,t)=f(x)$ for some f and consequently $f(x)=0$ since $u(x,0)=0$. But then, for $0<t<T-\varepsilon$, $u_x(1,t)=0=L\phi(u(1,t))>0$, a contradiction. If $v(1,t_0)=0$ for some t_0, $0<t_0\le T-\varepsilon$, $v_x(1,t_0)=0$ also. Therefore, by the second form of the strong maximum principle, [7, p. 190], $v\equiv 0$ in $\{(x,t)|0\le x<1, 0\le t\le t_0$ or $x=1$ and $0<t\le t_0\}$. Thus, as before, $u(x,t)=f(x)\equiv 0$ in this latter point set and hence $u_x(1,t_0)=0=L\phi(u(1,t_0))>0$. Therefore $v=u_t>0$ whenever it is defined except along $x=0$.

Remark 2.1. The content of Lemmas 2.2 and 2.3 is that the maximum of u in any closed domain $\bar{D}_T$ must occur at the point $(1,T)$.

COROLLARY 2.4. *The solution of problem* (α) *is unique.*

Proof. One lets u_1, u_2 be two solutions. If $w=e^{\lambda x-\lambda^2 t}(u_1-u_2)$, then w satisfies the same initial boundary value problem as the w of the preceding lemma except that $w_x=(\lambda+L\phi'(\xi))w$ where ξ is between $u_1(1,t)$ and $u_2(1,t)$. Since $w\ge 0$ and $-w\ge 0$ by the first part of the proof, we have $w\equiv 0$.

LEMMA 2.5. *Let* $f(x)=ax$ *where* $a<1$ *is a root of* $a=L\phi(a)$ *and let* u *solve problem* (α). *Then* $u(x,t)<f(x)$ *for all* $(x,t)\in D_T$.

Proof. Put $v(x,t)=f(x)-u(x,t)$. Then $v(0,t)=0, v(x,0)=f(x)>0$, $v_t=v_{xx}$ and $v_x(1,t)=a-L\phi(u(1,t))=L(\phi(a)-\phi(u(1,t)))=L\phi'(\xi)v$ where ξ is between a and $u(1,t)$. By the same argument with $w=e^{\lambda x-\lambda^2 t}v$ as in Lemma 2.3 we conclude that $v>0$ in D_T.

Remark 2.2. The equation $a=L\phi(a)$ need not have any solutions in $(0,1)$.

We shall assume for the purposes of this section that if $u\leq 1-\delta$ on $\bar{D}_T$, then u may be extended to be a larger domain $\bar{D}_{T+\sigma}$ on which $u\leq 1-\delta'$ $(\delta'<\delta)$ for some $\sigma>0$ and sufficiently small. This will be established later.

THEOREM 2.5. *Either* (a) u *exists on* D *and* $\lim_{t\to+\infty}u(x,t)=ax$ *where* $a=L\phi(a)$ *and* $a<1$ *or* (b) *for some* $T<\infty$, $\lim_{t\to T^-}u(1,t)=1$ (*u quenches in finite time*).

Proof. Suppose (b) fails. Then since $u(1,T)\geq u(x,t)$ on D_T for all T, by the comment on continuation, we may assume u exists (and is less than one) for all $t\geq 0$. Let

$$G(x,y)=\begin{cases}x, & 0\leq x\leq y\leq 1,\\ y, & 0\leq y\leq x\leq 1\end{cases}$$

denote the Green's function for d^2/dx^2 with boundary conditions $G(0,y)=G_x(1,y)=0$. By Lemma 2.3,

$$\lim_{t\to+\infty}u(x,t)=h(x)\qquad(\leq 1)$$

exists. Put

$$F(x,t)=\int_0^1 G(x,y)u(y,t)\,dy.$$

Then

$$F_t(x,t)=\int_0^1 G(x,y)u_t(y,t)\,dy>0$$

by Lemma 2.3. Using the differential equation and integrating by parts we find

$$F_t(x,t)=(yu_y-u)\Big|_0^x+xu_y\Big|_x^1=-u(x,t)+xL\phi(u(1,t)).$$

Clearly,

$$\lim_{t\to+\infty}F_t(x,t)=\begin{cases}-h(x)+xL\phi(h(1))\equiv M(x) & \text{if } h(1)<1,\\ +\infty & \text{if } h(1)=1.\end{cases}\tag{2.1}$$

where $M(x)\geq 0$. Now for any x,t we have

$$F(x,t)\leq\int_0^1 G(x,y)\,dy\leq\frac{1}{2}.$$

It is easy to see that if $f(t)$ is a bounded function such that $f'(t)\geq 0$ and $\lim_{t\to\infty}f'(t)=\alpha\geq 0$, then $\alpha=0$. Therefore $h(1)<1$ and $M(x)\equiv 0$ so that $h(x)=xL\phi(h(1))$ so that with $a=h(1)$ we have the theorem.

COROLLARY 2.6. *If* $a=L\phi(a)$ *has no solutions in* $(0,1)$ *then* $u(1,t)$ *reaches one in finite time.*

Example 2.1. $\phi(u)=(1-u)^{-\beta}$, $\beta>0$. Then it is easily checked that $a=L\phi(a)$ has no solutions if $L>L_0=\beta^\beta(1+\beta)^{-(1+\beta)}$, one solution smaller than one if $L=L_0$ and two solutions smaller than one if $L<L_0$. In particular, if $\beta=1$ and $0<L\leq\frac{1}{4}$

$$\lim_{t\to+\infty}u(x,t)=a_-x,$$

where $a_-=\frac{1}{2}(1-\sqrt{1-4L})$.

COROLLARY 2.7. *Suppose* L *is so large that* $u(1,t)$ *reaches one in finite time. Then* $u_t(1,t)$ *becomes infinite in finite time.*

Proof. We invoke (4.3) of this paper which is used to establish local existence. With $f(x)\equiv 0$ there we find that

$$u_t(x,t)=LG(x,1;t)\phi(u(1,t)) + L\int_0^T G(x,1;t-\eta)\phi'(u(1,\eta))u_\eta(1,\eta))\,d\eta.$$

Since $\phi'>0$ and $u_\eta\geq 0$ it follows that

$$u_t(1,t)\geq LG(1,1;t)\phi(u(1,t)),$$

where G is the Green's function following (4.1). Since $G(1,1,t)>0$ on $[0,\infty)$ and $\phi(u(1,t))\to+\infty$ in finite time, the result follows.

3. The hyperbolic problem (β). Here we consider weak solutions of (β).

DEFINITION 3.1. A continuous function u on $D_T\cup\Gamma_T$ is a *weak solution* of problem (β) if

(i) $u(1,t)<1$ for $0\leq t<T$;

(ii) $u(0,t)=u(x,0)=u_t(x,0)=0$, $0\leq x\leq 1$, $0\leq t<T$;

(iii) u has weak derivatives u_x, u_t, which, as functions of x are in $L^2(0,1)$ for each $t\in(0,T)$;

(iv) for every $\psi\in C_p^1(\bar{D}_T)$ with $\psi(0,t)=0$

$$\int_0^1\psi(y,t)u_t(y,t)\,dy=L\int_0^t\psi(1,\eta)\phi(u(1,\eta))\,d\eta + \int_0^t\int_0^1\left[\psi_\eta(y,\eta)u_\eta(y,\eta)-\psi_y(y,\eta)u_y(y,\eta)\right]dy\,d\eta \tag{3.1}$$

(C_p^1 denotes piecewise C^1 functions);

(v) The following conservation law holds:

$$E(t)\equiv\frac{1}{2}\int_0^1 u_t^2(x,t)\,dx+\frac{1}{2}\int_0^1 u_x^2(x,t)\,dx-L\int_0^{u(1,t)}\phi(\eta)\,d\eta=E(0)(=0). \tag{3.2}$$

Remark 3.1. Notice that the boundary condition at $x=1$ has been incorporated into (3.1) and (3.2) because $u_x(1,t)$ need not be defined for specific points. Notice also that (3.2) implies $0\leq u(1,t)$. Equation (3.2) can be obtained formally from (β) in the usual manner. See also Theorem 4.2.

We shall assume for the purposes of this section that if u is a weak solution on $D_T\cup\Gamma_T$ and $u\leq 1-\delta$ on $\{1\}\times[0,T]$ then u may be continued as a weak solution on $D_{(T+\sigma)}\cup\Gamma_{(T+\sigma)}$ for σ sufficiently small and positive while $u\leq 1-\delta'$ ($\delta'<\delta$) on $x=1$. This will be established in the next section.

THEOREM 3.1. *Let u be a weak solution on (β) on $D_T\cup\Gamma_T$. Let*

$$L_1=\sup_{0\leq\delta\leq 1}\Psi(\delta),$$

where

$$\Psi(\delta)=\frac{1}{2}(1-\delta)^2\left(\int_0^{1-\delta}\phi(\sigma)\,d\sigma\right)^{-1},$$

while

$$L_2=\sup_{0\leq\delta\leq 1}(1-\delta)/\phi(1-\delta).$$

(a) *If* $L<L_1$ *then* $T=+\infty$ *and* u *cannot quench, even in infinite time, i.e.,* $|u(x,t)|\leq 1-\delta$ *on* $x=1$ *for some* $\delta>0$.

(b) *If* $L>L_2$ *and* $\int_0^1\phi(\eta)\,d\eta<+\infty$, *then* $T<\infty$ *and* $\lim_{t\to T^-}u(1,t)=1$.

(c) *If* $L>L_2$ *and* $\int_0^1\phi(\eta)\,d\eta=+\infty$, *then* $T\leq\infty$ *and* $\lim_{t\to T^-}u(1,t)=1$.

Proof. We first note that since $\Psi(1)=0$ and $\Psi(0)\geq 0$, $L_1=\Psi(\delta_0)$ for some $\delta_0\in[0,1)$ which solves $\Psi'(\delta_0)=0$ or

$$\int_0^{1-\delta_0}\phi(\sigma)\,d\sigma=\frac{1}{2}(1-\delta_0)\phi(1-\delta_0)$$

($\Psi(1)=0$ by l'Hôpital's rule, $\Psi(0)>0$ if $\int_0^1\phi(\sigma)\,d\sigma<+\infty$, otherwise $\Psi(0)=0$). Therefore

$$L_1=\frac{(1-\delta_0)}{\phi(1-\delta_0)}\leq L_2,$$

as should be the case.

(a). For any $L<L_1$ there is a $\delta_1\in(0,1)$ such that $L<\Psi(\delta_1)$. Let T be the largest time such that $u(x,t)\leq 1-\delta_1$ on $x=1$. From (3.2) and the (sharp) inequality

$$u^2(x,t)\leq\int_0^1 u_x^2(x,t)\,dx,$$

we have, for $0<x<1$,

$$u^2(x,t)\leq 2L\int_0^{u(1,t)}\phi(\eta)\,d\eta. \tag{3.3}$$

If we take note of the monotonicity of the integral with respect to the upper limit, and take the supremum over $\{1\}\times[0,T]$ on the left of (3.3), we find that

$$L\geq\Psi(\delta_1),$$

which contradicts the choice of δ_1. This, together with the remarks on continuation, proves that $u(1,t)<1-\delta_1$. Using (3.3) and the definition of δ_1, (a) follows.

(b). Suppose (b) fails, i.e., $T=+\infty$ and $u<1$ on $x=1$. Define

$$F(x,t)=\int_0^1 G(x,y)u(y,t)\,dy,$$

where G is the Green's function given in Theorem 2.5. There results

$$F_t(x,t)=\int_0^1 G(x,y)u_t(y,t)\,dy.$$

Since for each x, $G(x,\cdot)$ is admissible in (3.1), we find

$$F_t(x,t)=L\int_0^t G(x,1)\phi(u(1,\eta))\,d\eta-\int_0^t\int_0^1 G_y(x,y)u_y(y,\eta)\,dy\,d\eta.$$

Therefore F_{tt} exists and

$$\begin{aligned}F_{tt}(x,t)&=Lx\phi(u(1,t))-\int_0^1 G_y(x,y)u_y(y,t)\,dy\\&=Lx\phi(u(1,t))-\int_0^x u_y(y,t)\,dy\\&=Lx\phi(u(1,t))-u(x,t).\end{aligned}$$

Since $0\leq u(1,t)<1$ and $L>L_2$, we have $L\geq u(1,t)/\phi(u(1,t))+\varepsilon$ for some $\varepsilon>0$. Thus $F_{tt}(1,t)\geq\varepsilon\phi(u(1,t))\geq\varepsilon\phi(0)$ so that $F(1,t)\geq\frac{1}{2}\varepsilon\phi(0)t^2$. On the other hand, if we take

square roots of both sides of (3.3) multiply through the resulting inequality by $G(1,y)$ and integrate over $[0,1]$, we find

$$\frac{1}{2}\varepsilon\phi(0)t^2 \le F(1,t) \le \sqrt{2L}\left(\int_0^1 G(1,y)\,dy\right)\left(\int_0^{u(1,t)}\phi(\eta)\,d\eta\right)^{1/2}. \tag{3.4}$$

Since we are assuming $\int_0^1 \phi(\sigma)\,d\sigma<\infty$, (3.4) is not possible for all $t>0$. Hence $u(1,t)$ reaches one in finite time.

(c) This also follows from (3.4). If $u(1,t)\le 1-\delta$ on $[0,\infty)$, then (3.4) will again be contradicted. Hence $\lim_{t\to T^-} u(1,t)=1$ where $T\le+\infty$.

Before stating an important corollary, we look at an example.

Example 3.2. If $\phi(u)=(1-u)^{-\beta}$, $\beta>0$, then u reaches one along $x=1$ in finite or infinite time provided

$$L>L_2(\beta)=\beta^{\beta}(1+\beta)^{-(1+\beta)}.$$

If $0<\beta<1$, part (b) of Theorem 3.1 applies. On the other hand, if $L<L_1(\beta)$, where

$$L_1(\beta)=\begin{cases}\max\limits_{0\le\delta\le 1}\dfrac{1}{2}(1-\delta)^2/\ln(1/\delta)=\delta_0(1-\delta_0), & \beta=1,\\[2ex] \max\limits_{0\le\delta\le 1}\dfrac{1}{2}(1-\beta)(1-\delta)^2/(1-\delta^{1-\beta})=\delta_0(1-\delta_0)^{\beta}, & \beta\ne 1,\end{cases}$$

where $2\ln\delta_0=1-1/\delta_0$ if $\beta=1$ and $2\delta_0^{\beta}-(1+\beta)\delta_0-(1-\beta)=0$ if $\beta\ne 1$, then u cannot quench, even in infinite time. To see this, one simply calculates $\Psi'(\delta)$ and shows that it has a unique root $\delta_0\in(0,1)$ while $\Psi'(\delta)$ changes from being positive for $\delta<\delta_0$ to being negative for $\delta>\delta_0$. We notice that for $\beta=1$, $L_2(\beta)=0.25$ while $L_1(\beta)\cong 0.20365$ and $\delta_0\simeq 0.2847$.

Remark 3.3. Since $\chi(\delta)\equiv(1-\delta)/\phi(1-\delta)$ vanishes at $\delta=0$ and $\delta=1$, $L_1=L_2$ provided there is $\delta_0\in(0,1)$ such that δ_0 maximizes both χ and Ψ. Then we have $\chi(\delta_0)=\Psi(\delta_0)$. This reduces to the requirement that the equations

$$\int_0^{1-\delta}\phi(\sigma)\,d\sigma=\frac{1}{2}(1-\delta)\phi(1-\delta), \qquad \phi(1-\delta)=(1-\delta)\phi'(1-\delta)$$

have a common solution $\delta=\delta_0$. This cannot happen if ϕ' is strictly increasing since then

$$\frac{(1-\delta_0)^2}{2}\phi'(1-\delta_0)=\int_0^{1-\delta_0}\eta\phi'(\eta)\,d\eta<\frac{(1-\delta_0)^2}{2}\phi'(1-\delta_0)$$

as an integration by parts shows. Thus these techniques are unlikely to yield optimal results.

COROLLARY 3.4. *If u solves (β) and*

(a) *$u(1,t)$ reaches one in finite time T, then*

$$\lim_{t\to T^-} u_x(1,t)=\lim_{t\to T^-} u_t(1,t)=+\infty, \quad \textit{or}$$

(b) *$u(1,t)$ reaches one in infinite time, then*

$$\limsup_{t\to+\infty}\max_{0\le x\le 1} u(x,t)=\limsup_{t\to+\infty}\int_0^1 u_x^2(1,t)\,dx=+\infty.$$

The proof of this corollary is postponed to the end of §4 because it depends upon a certain auxiliary function introduced in the next section.

COROLLARY 3.5. *If* $u(1,t)\le 1-\delta$ *for all* t, *then for all* $t>0$ *and all* x, $0<x<1$,

$$|u(x,t)|\le\sqrt{2L}\left(\int_0^{1-\delta}\varphi(\eta)\,d\eta\right).$$

Proof. This is an obvious consequence of (3.3).

4. Local existence. In this section we examine the questions of local existence and continuation of local solutions of problems (α), (β). Since these questions reduce to the study of nonlinear Volterra integral equations, copious detail will not be needed.

We begin with problem (α). The solution of

$$(\alpha_1)\qquad \begin{aligned} &w_t=w_{xx}+F(x,t), \qquad 0<x<1, \quad t>0,\\ &w(0,t)=0,\\ &w_x(1,t)=0\\ &w(x,0)=f(x) \end{aligned}$$

can be found by elementary means. It is

$$(4.1)\qquad w(x,t)=\int_0^1 G(x,y;t)f(y)\,dy+\int_0^t\int_0^1 G(x,y;t-\eta)F(y,\eta)\,dy\,d\eta,$$

where $G(x,y;t)$ is the heat kernel for the homogeneous problem. In fact, with $\lambda_n=\frac{1}{2}(2n+1)\pi$,

$$G(x,y;t)=2\sum_{n=1}^{\infty}\sin\lambda_n x\sin\lambda_n y\exp(-\lambda_n^2 t).$$

It is well known that $G>0$ on the half strip and $G_{xx}=G_{yy}=G_t$, $G_y(x,1;t)=G_x(1,y;t)=G(0,y;t)=G(x,0,t)=0$, and

$$\int_0^1 G(x,y;t)\,dy\le 1.$$

Consider next problem (α) *with inhomogeneous nonnegative* initial data $u(x,0)=f(x)$. If we set

$$w(x,t)=u(x,t)-xL\phi(u(1,t)),$$

then w solves problem (α_1) with $w(x,0)=f(x)-L\phi(u(1,0))$, $F(x,t)=-xL\phi'(u(1,t))u_t(1,t)$. Therefore u must solve, on $D_T\cup\Gamma_T$, the nonlinear Volterra equation

(4.2)

$$\begin{aligned} u(x,t)=xL\phi(u(1,t))&+\int_0^1 G(x,y;t)f(y)\,dy\\ &-L\phi(u(1,0))\int_0^1 yG(x,y;t)\,dy-L\int_0^t\int_0^1 G(x,y;t-\eta)\frac{d}{d\eta}\phi(u(1,\eta))\,dy\,d\eta. \end{aligned}$$

If one integrates by parts in this last integral, uses $G_t=G_{yy}$ and again integrates by parts, one sees that (4.2) takes the more pleasant form

$$(4.3)\qquad u(x,t)=\int_0^1 G(x,y;t)f(y)\,dy+L\int_0^t G(x,1;t-\eta)\phi(u(1,\eta))\,d\eta.$$

It is now a straightforward matter to prove the convergence (pointwise) of the following iteration scheme on $\bar{D}_T$ provided T is sufficiently small:

$$u_0(x,t)\equiv 0,$$
$$u_{n+1}(x,t)=\int_0^1 G(x,y;t)f(y)\,dy+L\int_0^t G(x,1;t-\eta)\phi(u_n(1,\eta))\,d\eta.$$

Since $G>0$ and ϕ is increasing, $u_n>0$ for all n. Moreover $u_1>u_0=0$ so $\phi(u_n(1,t))\geq \phi(u_{n-1}(1,t))$ if $u_n\geq u_{n-1}$. Thus by induction $u_{n+1}\geq u_n$ on $\bar{D}_T$. Now suppose $f(x)\leq 1-2\delta$ and $u_n\leq 1-\delta$, then $u_{n+1}\leq 1-\delta$ also provided T is so small that

$$(1-2\delta)+L\phi(1-\delta)\int_0^T G(x,1,T-\eta)\,d\eta\leq 1-\delta,$$

i.e., provided T is so small that

$$\sup_{0\leq x\leq 1}\int_0^T G(x,1,T-\eta)\,d\eta\leq \frac{\delta L^{-1}}{\phi(1-\delta)}.$$

Clearly this is always possible. Thus the sequence $\{u_n\}_{n=1}^{\infty}$ of iterates is an increasing sequence of continuous functions, bounded above by $1-\delta$ if T is sufficiently small. Now by the monotone convergence theorem, $\lim_{n\to\infty}u_n=u$ exists and satisfies (4.3) and hence (4.1). Thus we have established the following:

THEOREM 4.1. *If T is sufficiently small, then problem (α) possesses a unique solution, C^1 in t, C^2 in x in D_T and continuous in $\bar{D}_T$. Moreover, if $u\leq 1-\delta$ on $\bar{D}_T$ then u may be continued as a solution on $D_{T+T'}\cup\Gamma_{T+T'}$ for T' sufficiently small and $u\leq 1-\delta'$ where $\delta'<\delta$ on $\bar{D}_{T+T'}$.*

We turn next to problem (β). We consider first the inhomogeneous problem

(β_1)
$$w_{tt}=w_{xx}+F(x,t),\qquad 0<x<1,\quad t>0,$$
$$w(x,0)=f(x),$$
$$w_t(x,0)=g(x)$$
$$w(0,t)=w_x(1,t)=0,\qquad t>0.$$

We let

$$B=\{f\colon R^1\to R^1 \mid f,f' \text{ are piecewise continuous,}$$
$$f(x)=f(2-x)=-f(-x)=f(x+4)\}.$$

We extend f,g and $F(\cdot,t)$ for each t so that $f,g,F(\cdot,t)\in B$. This amounts to requiring that $f(0)=g(0)=F(0,t)=0$ and that f,f',g,g',F,F_x are continuous on $[0,1]$. The solution of (β_1), which is then given by the d'Alembert formula

$$u(x,t)=\frac{1}{2}[f(x+t)+f(x-t)]+\frac{1}{2}\int_{x-t}^{x+t}g(\sigma)\,d\sigma+\frac{1}{2}\int_0^t\int_{x-t+\eta}^{x+t-\eta}F(\xi,\eta)\,d\xi\,d\eta, \tag{4.4}$$

again has the property that $w(\cdot,t)\in B$ for each $t\geq 0$.

For the purpose of the argument that follows, we shall assume $\phi\in C^2(-\infty,1)$. Let $H(x)\in B$ be defined by

$$H(x)=\begin{cases}2-x, & 1<x<2,\\ x, & -1\leq x\leq 1,\\ -(2+x), & -2\leq x<-1\end{cases}$$

on $(-2,2)$.

Suppose we have a solution of (β) *with inhomogeneous initial data*, f, g. Define, for $0 \le x \le 1$,

$$w(x,t) = u(x,t) - LH(x)\phi(u(1,t)).$$

Then w solves (β_1) with

$$\begin{aligned} &F(x,t) = -LH(x)\, d^2\phi(u(1,t))/dt^2, \\ &w(x,0) = f(x) - LH(x)\phi(f(1)), \\ &w_t(x,0) = g(x) - LH(x)\phi'(f(1))g(1), \\ &w(0,t) = w_x(1,t) = 0. \end{aligned}$$

Therefore, for w, the data are in B (for each t). Therefore u solves problem (β) weakly if and only if u solves

$$\begin{aligned} (4.5) \qquad u(x,t) = u_0(x,t) &- \frac{1}{2}L\phi(f(1))[H(x+t)+H(x-t)] \\ &+ LH(x)\phi(u(1,t)) - \frac{1}{2}Lg(1)\phi'(f(1))\int_{x-t}^{x+t} H(\eta)\, d\eta \\ &- \frac{1}{2}L\int_0^t \frac{d^2}{d\eta^2}\phi(u(1,\eta))\left(\int_{x-t+\eta}^{x+t-\eta} H(\xi)\, d\xi\right) d\eta, \end{aligned}$$

where $u_0(x,t)$, the so-called free solution, is given by

$$(4.6) \qquad u_0(x,t) = \frac{1}{2}[f(x+t)+f(x-t)] + \frac{1}{2}\int_{x-t}^{x+t} g(\eta)\, d\eta.$$

If one integration by parts is carried out in the last integral on the right of (4.5), we find

$$\begin{aligned} (4.7) \quad u(x,t) = u_0(x,t) &+ LH(x)\phi(u(1,t)) - \frac{1}{2}L\phi(f(1))[H(x+t)+H(x-t)] \\ &- \frac{1}{2}L\int_0^t \frac{d}{d\eta}\phi(u(1,\eta))[H(x+t-\eta)+H(x-t+\eta)]\, d\eta. \end{aligned}$$

The quantity in brackets in the integral on the right of (4.7) is piecewise continuously differentiable. Thus we may integrate by parts one more time and obtain

$$(4.8) \quad u(x,t) = u_0(x,t) + \frac{1}{2}L\int_0^t \phi(u(1,\eta))\frac{d}{d\eta}[H(x-t+\eta)+H(x+t-\eta)]\, d\eta$$

where now $-\infty < x < \infty$, $t > 0$. The kernel in the integrand in (4.8) is piecewise constant and cannot exceed two in absolute value.

It is clear from (4.8) that if u solves (4.8) in $R^1 \times [0,\tau]$ and is continuous there, then u_x and u_t exist and are piecewise continuously differentiable except on the lines $x = n$, $x \pm t = n$ where n is an integer. In fact, the second derivatives exist except on that point set, so that the solution is classical except on that point set. Therefore (4.8) need only be solved in the larger space

$$\begin{aligned} B_\tau = \{u:\, &R^1 \times [0,\tau] \to R^1 | u(x,t) = u(1-x,t) \\ &= -u(-x,t) = u(x+4,t),\ u \text{ is continuous}\}. \end{aligned}$$

Let

$$B(\delta) = \{u \in B_\tau | |u(1,t)| \le 1-\delta\}.$$

For $u \in B_\tau$, let

$$\|u\| = \sup\{|u(x,t)|, 0 \le t \le \tau, 0 \le x \le 1\}.$$

Let, for $f, g \in B$, τ be so small that

$$|u_0(1,t)| = \left| \frac{1}{2}[f(1+t)+f(1-t)] + \frac{1}{2}\int_{1-t}^{1+t} g(\eta)\,d\eta \right| \le 1-2\delta \tag{4.9}$$

for $0 \le t \le \tau_1$ say. This can be accomplished if $|f(1)| < 1-2\delta$. Let

$$B(\rho,\delta,u_0) = \{u \in B(\delta) \mid \|u-u_0\| \le \rho\}.$$

Define

$$\mathcal{T}: B(\rho,\delta,u_0) \to B(\rho,\delta,u_0)$$

by

$$(\mathcal{T}u)(x,t) = u_0(x,t) + \frac{1}{2}L\int_0^t \phi(u(1,\eta))K(x,t,\eta)\,d\eta,$$

where

$$K(x,t,\eta) = \frac{d}{d\eta}[H(x+t-\eta)+H(x-t+\eta)]$$

almost everywhere (except where $x \pm (t-\eta)$ is an integer).

We need to show that $\mathcal{T}$ is well defined and that it is a contraction. We note that if $u \in B(\rho,\delta,u_0)$,

$$(\mathcal{T}u)(1,t) = u_0(1,t) + \frac{1}{2}L\int_0^t [H'(1-(t-\eta)) - H'(1+(t-\eta))]\phi(u(1,\eta))\,d\eta.$$

Therefore, since $|K| \le 2$,

$$-1+2\delta - L\tau\phi(1-\delta) \le (\mathcal{T}u)(1,t) \le 1-2\delta + L\tau\phi(1-\delta).$$

Thus if

$$\tau \le \tau_2 \equiv \delta/(L\phi(1-\delta)), \tag{4.10}$$

we have that $|\mathcal{T}u(1,t)| \le 1-\delta$ also. Moreover

$$\|\mathcal{T}u - u_0\| \le L\tau\phi(1-\delta) \le \rho$$

if

$$\tau \le \tau_3 \equiv \rho\tau_2/\delta. \tag{4.11}$$

Thus $\mathcal{T}$ is well defined. Also one verifies readily that

$$\|\mathcal{T}u - \mathcal{T}v\| \le L\tau\left[\sup_{|\xi| \le 1-\delta} \phi'(\xi)\right]\|u-v\|$$

if $u, v \in \mathcal{B}(\rho,\delta,u_0)$. Thus $\mathcal{T}$ will be a contraction if, in addition to (4.9)–(4.11),

$$\tau < \tau_4 \equiv 1/\left[L \sup_{|\xi| \le 1-\delta} \phi'(\xi)\right]. \tag{4.12}$$

(This assumes $\phi' \not\equiv 0$ in any neighborhood of zero.) Therefore for $T<\min(\tau_1,\tau_2,\tau_3,\tau_4)$ we have proved the following theorem:

THEOREM 4.2. *There exists, for any $L>0$ and $\delta\in(0,1)$ a unique weak solution of problem (β) on some domain $\overline{D}_T$ for $T=T(L,\delta)$ sufficiently small, which satisfies $|u(1,t)| \le 1-\delta$ for $0\le t\le T$. This solution is classical except on the characteristics, so that* (3.1) *and* (3.2) *hold and therefore $u(1,t)\ge 0$ on $[0,T]$. Moreover if $\delta'<\delta$, this solution may be continued to $D_{T+T'}$ with $0\le u(1,t)\le 1-\delta'$ on $[0,T+T']$ for sufficiently small $T'>0$. The extended (in x) solution satisfies* (4.8) *on $R^1\times[0,T]$.*

Proof of Corollary 3.4. (a) From (4.8) with $f=g=0$, we see that

$$u(1,t)=\frac{1}{2}L\int_0^t \phi(u(1,\eta))\left[H'(1-(t-\eta))-H'(1+(t-\eta))\right]dy. \tag{4.13}$$

A few moments reflection will convince the reader that

$$H'(1-x)-H'(1+x)=2H'(x-1).$$

Using this in (4.13) and taking the (distribution) derivative of the result yields (where n is the largest integer such that $2n\le t-1$),

$$u_t(1,t)=L\phi(u(1,t))+2L\sum_{p=0}^{n}(-1)^p\phi(u(1,t-2p-1)). \tag{4.14}$$

If u quenches in (finite) time T, and if N is the largest integer such that $2N\le T-1$, then as $t\to T^-$ the sum on the right of (4.14) approaches

$$\sum_{p=0}^{N}(-1)^p\phi(u(1,T-2p-1))$$

while the first term becomes unbounded. This proves part (a) of the corollary.

To prove part (b), we see from part (b) of the proof of Theorem 3.1 that if $u(1,t)$ reaches one in infinite time then

$$\lim_{t\to+\infty}\int_0^1 G(1,y)u(y,t)\,dy=+\infty.$$

Thus there is a sequence of points (x_n,t_n) with $0<x_n<1$ and $t_n\to+\infty$ such that

$$\lim_{n\to\infty}u(x_n,L_n)=+\infty.$$

Using these points in the inequality preceding (3.3) we find

$$\lim_{n\to\infty}\int_0^1 u_x^2(x,t_n)\,dx=+\infty.$$

Acknowledgment. The author thanks Professor Gary Lieberman for some helpful discussions and the referees for several comments which improved the earlier version of this paper.

REFERENCES

[1] A. ACKER AND W. WALTER, *On the global existence of solutions of parabolic differential equations with a singular nonlinear term*, Nonlinear Analysis, 2 (1978), pp. 499–505.

[2] _____, *The quenching problem for nonlinear parabolic equations*, Lecture Notes in Mathematics 564, Springer-Verlag, New York, 1976.

[3] P. H. CHANG AND H. A. LEVINE, *The quenching of solutions of semilinear hyperbolic equations*, this Journal, 12 (1981), pp. 893–903.

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

[5] H. KAWARADA, *On the solutions of initial boundary value problems for* $u_t = u_{xx} + 1/(1-u)$, Publ. RIMS, Kyoto Univ., 10 (1975), pp. 729–736.

[6] H. A. LEVINE AND J. T. MONTGOMERY, *The quenching of solutions of some nonlinear parabolic equations*, this Journal, 11 (1980), pp. 842–847.

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

[8] W. WALTER, *Differential and Integral Inequalities*, Ergib. der Math. Band 55, Springer-Verlag, Berlin, 1970.

[9] J. J. STOKER, *Nonlinear Vibrations*, Interscience, New York, 1950.

[10] N. MINORSKY, *Introduction to Nonlinear Mechanics*, J. W. Edwards, Ann Arbor, MI, 1947.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0007 $01.25/0

ISOPERIMETRIC INEQUALITIES IN THE TORSION AND CLAMPED MEMBRANE PROBLEMS FOR CONVEX PLANE DOMAINS*

L. E. PAYNE† AND G. A. PHILIPPIN‡

Abstract. Bounds for the curvature of the level curve of the torsion function through an arbitrary point in a convex region D, are used to derive improved isoperimetric inequalities for maximum stress, the torsional rigidity and other functionals. These inequalities are exact if D is either a circle or an infinite strip. A similar procedure is used in the clamped membrane problem, and again improved isoperimetric inequalities are derived.

1. Introduction. In this paper we make use of some recent results of Makar-Limanov [4] and of Acker, Payne and Philippin [1] to sharpen a number of inequalities that have appeared in the recent literature. In [4] it was shown that the level curves of the torsion function for a convex plane region D are convex. The analogous result for the first eigenfunction in the fixed membrane problem was established by Brascamp and Lieb [2]. We show in this paper that the proof of Makar-Limanov actually leads to pointwise bounds for the curvature of the level curve of the torsion function through an arbitrary point in D. The method employed by Brascamp and Lieb [2] does not yield the corresponding bounds for the curvature of the level curves in the clamped membrane problem; however, the method of proof employed in [1] does yield such bounds. It is these bounds for the curvature of level curves which enable us to improve a number of inequalities derived in [5], [6], and [8].

In §2, we consider the elastic torsion problem and derive isoperimetric inequalities for the maximum values of the torsion function ψ and its gradient, as well as isoperimetric inequalities for the torsional rigidity. These inequalities have the unique feature that the equality sign holds not only when D is the interior of a circle but also in the limit as D tends to the infinite strip. In §3, we derive improved bounds for the absolute value of the gradient of the first eigenfunction u in the fixed membrane problem, as well as a sharper bound for the corresponding eigenvalue λ_1. To make these latter estimates explicit, it would be helpful to have a nonzero lower bound for the minimum value of $|\operatorname{grad} u|$ on ∂D, but the authors are not aware of the existence of any such bound in the literature. Section 4 gives an improvement of the lower bound for $\lambda_1\psi_{\max}$ derived in [6].

2. The torsion problem. In this section, we consider the problem of the torsional rigidity of a beam whose cross section is a convex bounded plane domain D,

$$\Delta\psi=-2 \quad \text{in } D, \qquad \psi=0 \quad \text{on } \partial D. \tag{2.1}$$

It is well known that all level curves of the stress function $\psi(x)$ are also convex. The proof of this statement has been given by Makar-Limanov in [4]. It is based on the fact that the function

$$M(x)=\psi_{,ij}\psi_{,i}\psi_{,j}-\sigma^2\Delta\psi+\psi\left[(\Delta\psi)^2-\psi_{,ij}\psi_{,ij}\right], \tag{2.2}$$

* Received by the editors January 15, 1982.

† Cornell University, Ithaca, New York, 14853. The research of this author was supported partly by the National Science Foundation under grant NSF MCS 79-19358.

‡ Université Laval, Québec, G1K 7P4, Canada. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

with

$$\sigma^2=\psi_{,i}\psi_{,i}, \tag{2.3}$$

is superharmonic, which implies that $M(x)$ takes its minimum value on the boundary ∂D of the domain D:

$$M(x)\geq\min_{\partial D}M(x)=M_{\min}=\min_{\partial D}(k\sigma^3)\geq 0. \tag{2.4}$$

We shall use this fact to obtain upper and lower bounds for the curvature

$$k=-\left(\frac{\psi_{,i}}{\sigma}\right)_{,i} \tag{2.5}$$

of the level curves $\psi=$const. from which various isoperimetric inequalities can be established. It turns out that these inequalities will be sharper then those given by Payne in [5]. Using normal coordinates with respect to the level curve $\psi=$const., we have

$$\sigma^2=\psi_{,i}\psi_{,i}=\psi_n^2+\psi_s^2, \tag{2.6}$$

$$\Delta\psi=\psi_{nn}+k\psi_n, \tag{2.7}$$

$$\psi_{,ij}\psi_{,ij}=2\psi_{sn}^2+k^2\psi_n^2+\psi_{nn}^2, \tag{2.8}$$

where $\psi_n=\partial\psi/\partial n$ is the outward normal derivative, and $\psi_s=\partial\psi/\partial s$ is the tangential derivative of ψ along the level curve. Thus (2.4) can be rewritten as

$$M(x)=\sigma^3k-2\psi\{2k\psi_n+k^2\psi_n^2+\psi_{ns}^2\}\geq M_{\min}\geq 0, \tag{2.9}$$

which is equivalent to the following quadratic inequality for the curvature:

$$k^2-\frac{k}{2\psi\sigma}(\sigma^2+4\psi)+\frac{\psi_{ns}^2}{\sigma^2}+\frac{M_{\min}}{2\psi\sigma^2}\leq 0, \tag{2.10}$$

or

$$\left(k-\frac{\Phi}{4\psi\sigma}\right)^2\leq\left(\frac{\Phi}{4\psi\sigma}\right)^2-\frac{\psi_{ns}^2}{\sigma^2}-\frac{M_{\min}}{2\psi\sigma^2}\leq\left(\frac{\Phi}{4\psi\sigma}\right)^2-\frac{M_{\min}}{2\psi\sigma^2}. \tag{2.11}$$

Here we have used the abbreviation

$$\Phi(x)=\sigma^2+4\psi. \tag{2.12}$$

Inequality (2.11) leads to the following inequalities for the curvature k of the level curve in terms of the stress function ψ, the stress $\sigma=\sqrt{\psi_{,i}\psi_{,i}}$ and $M_{\min}$:

$$\frac{\Phi}{4\psi}\left\{1-\sqrt{1-\frac{M_{\min}}{2\psi\sigma^2}\left(\frac{4\psi\sigma}{\Phi}\right)^2}\right\}\leq k\sigma\leq\frac{\Phi}{4\psi}\left\{1+\sqrt{1-\frac{M_{\min}}{2\psi\sigma^2}\left(\frac{4\psi\sigma}{\Phi}\right)^2}\right\}. \tag{2.13}$$

On the other hand we have

$$\frac{\partial\Phi}{\partial n}=-2k\sigma^2, \tag{2.14}$$

so that we obtain differential inequalities relating Φ and its normal derivative by multiplying (2.13) by -2σ:

$$(2.15)\quad -\frac{\Phi\sigma}{2\psi}\left\{1+\sqrt{1-\frac{M_{\min}}{2\psi\sigma^2}\left(\frac{4\psi\sigma}{\Phi}\right)^2}\right\}\leq\frac{\partial\Phi}{\partial n}\leq-\frac{\Phi\sigma}{2\psi}\left\{1-\sqrt{1-\frac{M_{\min}}{2\psi\sigma^2}\left(\frac{4\psi\sigma}{\Phi}\right)^2}\right\}.$$

These inequalities may be expressed in a more compact form if we replace Φ by a new function θ, defined as

$$(2.16)\quad \theta=\frac{\Phi}{\sqrt{\psi}}=\frac{\sigma^2}{\sqrt{\psi}}+4\sqrt{\psi}\,.$$

Differentiation of (2.16) yields

$$(2.17)\quad \frac{\partial\Phi}{\partial n}=\sqrt{\psi}\,\frac{\partial\theta}{\partial n}-\frac{\theta\sigma}{2\sqrt{\psi}},$$

and insertion of (2.16) and (2.17) into (2.15) leads to the inequalities

$$(2.18)\quad \frac{\partial\psi}{\partial n}\sqrt{\left(\frac{\theta}{2}\right)^2-2M_{\min}}\leq\psi\,\frac{\partial\theta}{\partial n}\leq-\frac{\partial\psi}{\partial n}\sqrt{\left(\frac{\theta}{2}\right)^2-2M_{\min}}\,.$$

Inequalities (2.18) state that the differentials $d\psi$ and $d\theta$ of the two functions $\psi(x)$, $\theta(x)$ along any orthogonal trajectory of $\psi=\text{const.}$ (also called *fall line* of ψ) in the inward direction are related as follows:

$$(2.19)\quad \frac{d\psi}{\psi}\geq\frac{d\theta}{\sqrt{\left(\frac{\theta}{2}\right)^2-2M_{\min}}}\geq-\frac{d\psi}{\psi}.$$

We now integrate (2.19) along the fall line from the point $P(x)\in D$ to the (unique) critical point of ψ, which gives

$$(2.20)\quad \sqrt{\frac{\psi}{\psi_{\max}}}\leq\frac{\theta+\sqrt{\theta^2-8M_{\min}}}{\theta_0+\sqrt{\theta_0^2-8M_{\min}}}\leq\sqrt{\frac{\psi_{\max}}{\psi}}\,,$$

with

$$(2.21)\quad \theta_0=\Phi\psi^{-1/2}\big|_{\psi=\psi_{\max}}=4\sqrt{\psi_{\max}}\,.$$

Multiplying the second inequality in (2.20) by $\sqrt{\psi}\,(\theta_0+\sqrt{\theta_0^2-8M_{\min}})$ and using again (2.16), we obtain

$$(2.22)\quad \sqrt{\Phi^2-8\psi M_{\min}}\leq\sqrt{\psi_{\max}}\left(\theta_0+\sqrt{\theta_0^2-8M_{\min}}\right)-\Phi.$$

Squaring (2.22) and solving for Φ leads to

$$(2.23)\quad \Phi\leq\frac{4\psi M_{\min}}{\sqrt{\psi_{\max}}\left(\theta_0+\sqrt{\theta_0^2-8M_{\min}}\right)}+\frac{1}{2}\sqrt{\psi_{\max}}\left(\theta_0+\sqrt{\theta_0^2-8M_{\min}}\right).$$

The first term on the right-hand side of (2.23) may be rewritten as follows:

$$\frac{4\psi M_{\min}}{\sqrt{\psi_{\max}}\,\theta_0\left(1+\sqrt{1-8M_{\min}/\theta_0^2}\right)}=2\psi\left[1-\sqrt{1-\frac{M_{\min}}{2\psi_{\max}}}\right], \tag{2.24}$$

so that (2.23) leads finally to the following upper bound for σ:

$$\sigma^2\leq\alpha(\psi_{\max}-\psi), \tag{2.25}$$

with

$$\alpha=2\left[1+\sqrt{1-\frac{M_{\min}}{2\psi_{\max}}}\right]. \tag{2.26}$$

In particular, the maximum value of σ which occurs on the boundary ∂D of D, is bounded above by

$$\sigma_{\max}^2\leq\alpha\psi_{\max}. \tag{2.27}$$

Inequality (2.25) is an improvement on an earlier result by Payne, who established in [5], that

$$\sigma^2\geq 4(\psi_{\max}-\psi). \tag{2.28}$$

Inequality (2.28) is isoperimetric in the limit as D degenerates into an infinite strip, whereas (2.25) is exact for the circle and for the infinite strip, since the function $M(x)$ introduced in (2.2) is a constant for any ellipse. For practical purposes, a lower bound for $M_{\min}$ is needed. Such a bound is easily obtained using the isoperimetric inequality

$$\min_{\partial D}\sigma=\sigma_{\min}\geq\frac{1}{k_{\max}}, \tag{2.29}$$

established by the authors in [7]. We have then

$$M_{\min}=\min_{\partial D}(k\sigma^3)\geq k_{\min}\sigma_{\min}^3\geq k_{\min}k_{\max}^{-3}. \tag{2.30}$$

With this bound for $M_{\min}$ in (2.26), the equality sign still holds in (2.25) for the circle and for the strip. Let us mention that (2.25) remains true for a nonconvex domain D with

$$\alpha=\alpha_0=2\left[1+\sqrt{1-\frac{M_{\min}}{2\psi_0}}\right], \tag{2.31}$$

where ψ_0 is the smallest value ψ can take on the set of its critical points. Note, however, that now $M_{\min}$ is negative.

A different bound for σ^2 has been obtained by Fu and Wheeler [3], i.e.,

$$\sigma^2\leq d(2-k_{\min}d), \tag{2.32}$$

where d is the radius of the largest circle inscribed in D. Here again the equality sign holds for the circle and for the infinite strip.

An upper bound for $\psi_{\max}$ may be obtained by integrating the inequality (2.25) along a ray from the point $P\in D$ to the nearest point P_0 on the boundary. If r is the distance from P, we have

$$-\frac{d\psi}{dr}\leq\sigma\leq\sqrt{\alpha(\psi_{\max}-\psi)} \tag{2.33}$$

from which we obtain

$$2\sqrt{\psi_{\max}}-2\sqrt{\psi_{\max}-\psi}\leq\sqrt{\alpha}\,d. \tag{2.34}$$

Evaluated at the critical point, (2.34) gives

$$\psi_{\max}\leq\frac{\alpha}{4}d^2\leq d^2\left(1-\frac{M_{\min}}{8\psi_{\max}}\right), \tag{2.35}$$

or

$$d^2-d\sqrt{d^2-\frac{M_{\min}}{2}}\leq 2\psi_{\max}\leq d^2+d\sqrt{d^2-\frac{M_{\min}}{2}}\,. \tag{2.36}$$

We mention finally that isoperimetric inequalities involving functionals defined on D, such as the torsional rigidity P given by

$$P=\iint_D|\operatorname{grad}\psi|^2dx=2\iint_D\psi\,dx, \tag{2.37}$$

may be obtained from (2.25) by integration over D or along the boundary ∂D. We have, for instance,

$$P\leq\frac{2\alpha}{\alpha+2}A\psi_{\max}, \tag{2.38}$$

$$\psi_{\max}\geq\frac{4A^2}{\alpha L^2}, \tag{2.39}$$

where A is the area of D and L is the length of ∂D.

3. The membrane problem. The method indicated in §2 can also be applied to the first eigenfunction $u(x)$ of a vibrating membrane defined in a bounded plane domain D and fixed on its boundary ∂D:

$$\Delta u+\lambda_1 u=0 \quad \text{in } D, \qquad u=0 \quad \text{on } \partial D. \tag{3.1}$$

In [1], Acker, Payne and Philippin introduced the function

$$\Pi(x)=\frac{2kq^3}{u}+(\Delta u)^2-u_{,ij}u_{,ij}, \qquad q=\sqrt{u_{,i}u_{,i}}, \tag{3.2}$$

to establish their version of the proof that if D is strictly convex, then the level curves of $u(x)$ are also convex. In contrast to $M(x)$, defined in (2.2), $\Pi(x)$ takes a positive minimum value at some interior point P of D. In fact we have

$$\Pi(x)\geq\Pi_{\min}=\frac{1}{2}\left(\frac{q^2+\lambda_1u^2}{u}\right)^2_{\min}>0 \quad \forall x\in D. \tag{3.3}$$

Using normal coordinates, (3.3) becomes, as shown in [1], a quadratic inequality for the curvature k of the level curves $u=\text{const.}$:

$$\left(k-\frac{\phi}{2uq}\right)^2\leq\frac{1}{4q^2}\left\{\left(\frac{\phi}{u}\right)^2-\left(\frac{\phi}{u}\right)^2_{\min}\right\}-\frac{u_{ns}^2}{q^2}, \tag{3.4}$$

with

$$\phi=q^2+\lambda_1u^2. \tag{3.5}$$

Solving (3.4) after dropping the last term we obtain the following bounds for the curvature:

$$\frac{\phi}{2u}\left\{1-\sqrt{1-\left(\frac{\phi}{u}\right)^2_{\min}\left(\frac{\phi}{u}\right)^{-2}}\right\}\leq kq\leq\frac{\phi}{2u}\left\{1+\sqrt{1-\left(\frac{\phi}{u}\right)^2_{\min}\left(\frac{\phi}{u}\right)^{-2}}\right\}. \tag{3.6}$$

We note now that if we take the normal derivative of ϕ along a level curve $u=\text{const.}$, we have

$$\frac{\partial\phi}{\partial n}=-2kq^2. \tag{3.7}$$

The inequalities (3.6) and (3.7) then lead to

$$q\left\{-\frac{\phi}{u}-\sqrt{\left(\frac{\phi}{u}\right)^2-\left(\frac{\phi}{u}\right)^2_{\min}}\right\}\leq\frac{\partial\phi}{\partial n}\leq q\left\{-\frac{\phi}{u}+\sqrt{\left(\frac{\phi}{u}\right)^2-\left(\frac{\phi}{u}\right)^2_{\min}}\right\}, \tag{3.8}$$

which can be reduced to a more compact form in terms of the new function $\Omega(x)$, defined as

$$\phi=u\Omega. \tag{3.9}$$

We have

$$\frac{\partial u}{\partial n}\sqrt{\Omega^2-\Omega^2_{\min}}\leq u\frac{\partial\Omega}{\partial n}\leq-\frac{\partial u}{\partial n}\sqrt{\Omega^2-\Omega^2_{\min}}, \tag{3.10}$$

with

$$\Omega_{\min}=\left(\frac{\phi}{u}\right)_{\min}. \tag{3.11}$$

An integration of inequalities (3.10) from a point $P\in D$ to the (unique) critical point of u along a fall line leads to

$$\frac{u}{u_{\max}}\leq\frac{\Omega+\sqrt{\Omega^2-\Omega^2_{\min}}}{\Omega_0+\sqrt{\Omega_0^2-\Omega^2_{\min}}}\leq\frac{u_{\max}}{u}, \tag{3.12}$$

with

$$\Omega_0=\frac{\phi}{u}\Big|_{u=u_{\max}}=\lambda_1 u_{\max}. \tag{3.13}$$

From the second inequality in (3.12) we obtain after some reduction

$$\phi\leq\frac{1}{2}\left(A+\frac{u^2\Omega^2_{\min}}{A}\right), \tag{3.14}$$

with

$$A=\left(\Omega_0+\sqrt{\Omega_0^2-\Omega^2_{\min}}\right)u_{\max}=\lambda_1 u^2_{\max}\left(1+\sqrt{1-\frac{\Omega^2_{\min}}{\Omega_0^2}}\right). \tag{3.15}$$

Inequality (3.14), whose last term may be rewritten as

$$\frac{u^2\Omega^2_{\min}}{A}=\lambda_1 u^2\left[1-\sqrt{1-\frac{\Omega^2_{\min}}{\Omega_0^2}}\right], \tag{3.16}$$

leads finally to the following upper bound for q^2:

$$q^2 \leq \beta\left(u_{\max}^2 - u^2\right), \tag{3.17}$$

with

$$\beta = \frac{\lambda_1}{2}\left(1 + \sqrt{1 - \frac{\Omega_{\min}^2}{\Omega_0^2}}\right). \tag{3.18}$$

Inequality (3.17) is valid for *strictly convex domains* only, in which case it is an improvement on an earlier result by Payne and Stakgold [8], who established that

$$q^2 \leq \lambda_1\left(u_{\max}^2 - u^2\right). \tag{3.19}$$

Let us remark that (3.19) is exact for the infinite strip, whereas inequality (3.17) is not directly applicable. For practical purposes a lower bound for $\Omega_{\min}$ is needed. Using the arithmetic geometric mean inequality we established in [1] that $\Omega(x)$ is bounded below by $q_{\min}\sqrt{\lambda_1}$ where $q_{\min}$ (>0) is the minimum value of q on the boundary ∂D. We have therefore

$$\beta \leq \frac{\lambda_1}{2}\left(1 + \sqrt{1 - \frac{q_{\min}^2}{\lambda_1 u_{\max}^2}}\right). \tag{3.20}$$

Integrating (3.17) from a point P in D to the nearest point $P_0 \in \partial D$, we obtain

$$\arcsin\left(\frac{u}{u_{\max}}\right) \leq \sqrt{\beta}\, d, \tag{3.21}$$

where d is the radius of the largest inscribed circle in D. Evaluated at the critical point, (3.21), together with (3.20), yields

$$\lambda_1 \geq \frac{1}{4d^2}\left(\pi^2 + \frac{q_{\min}^2 d^2}{u_{\max}^2}\right). \tag{3.22}$$

4. A further application. In this section we will establish the following result:

$$\psi_{\max} \geq \frac{\alpha}{4\lambda_1} j_{1-2/\alpha}^2, \tag{4.1}$$

where $j_{1-2/\alpha}$ is the first zero of the Bessel function of order $1-2/\alpha$. The other notation has been introduced in §§2 and 3. Inequality (4.1) is isoperimetric with equality if D is a circle or an infinite strip. It is a sharp version of a result established by Payne in [6].

For the proof, we introduce the auxiliary one-dimensional boundary value problem

$$\begin{aligned} &f''(t) + \frac{a}{t} f'(t) + f(t) = 0, \qquad t \in \left(0, \delta\sqrt{\psi_{\max}}\right), \\ &f'(0) = f\left(\delta\sqrt{\psi_{\max}}\right) = 0, \end{aligned} \tag{4.2}$$

which has the positive solution

$$f(t) = t^{(1-a)/2} J_{(1-a)/2}(t), \tag{4.3}$$

if the parameters a and δ are related as follows:

$$\delta = j_{(1-a)/2}\, \psi_{\max}^{-1/2}. \tag{4.4}$$

The inequality (4.1) is then a simple consequence of the fact that the positive function $h(x)$, defined in D as

$$h(x)=f\left(\delta\sqrt{\psi_{\max}-\psi(x)}\right), \tag{4.5}$$

satisfies the differential inequality

$$\Delta h+\lambda^* h\leq 0 \quad \text{in } D, \tag{4.6}$$

with

$$\lambda^*=\frac{\alpha j_{1-2/\alpha}^2}{4\psi_{\max}}, \tag{4.7}$$

and

$$h=0 \quad \text{on } \partial D. \tag{4.8}$$

Indeed we have with (4.6) and (4.8),

$$\begin{aligned} 0\geq&\iint_D\left[u(\Delta h+\lambda^* h)-h(\Delta u+\lambda_1 u)\right]dx \\ =&\iint_D uh(\lambda^*-\lambda_1)\,dx+\oint_{\partial D}\left(u\frac{\partial h}{\partial n}-h\frac{\partial u}{\partial n}\right)ds \\ =&\iint_D uh(\lambda^*-\lambda_1)\,dx, \end{aligned} \tag{4.9}$$

from which we easily conclude the desired result: $\lambda^*\leq\lambda_1$. To prove (4.6), (4.7), we differentiate (4.5):

$$h_{,k}=-\frac{f'\delta\psi_{,k}}{2\sqrt{\psi_{\max}-\psi}}, \tag{4.10}$$

where f' denotes the derivative of f with respect to its argument. Another differentiation leads to

$$\Delta h+\lambda^* h=-\frac{\delta f'}{\sqrt{\psi_{\max}-\psi}}\left[\frac{a+1}{4}\frac{\psi_{,k}\psi_{,k}}{\psi_{\max}-\psi}-1\right]-\lambda^* f\left[\frac{\delta^2}{4\lambda^*}\frac{\psi_{,k}\psi_{,k}}{\psi_{\max}-\psi}-1\right], \tag{4.11}$$

where we have eliminated f'' using (4.2). In view of (2.25) we select a and λ^* so that

$$\frac{a+1}{4}=\frac{\delta^2}{4\lambda^*}=\frac{1}{\alpha}, \tag{4.12}$$

where α is defined in (2.26). With this choice we have

$$\Delta h+\lambda^* h=\delta^2\left(\alpha-\frac{\psi_{,k}\psi_{,k}}{\psi_{\max}-\psi}\right)\left(\frac{f}{4}+\frac{1}{\alpha}\frac{f'}{\delta\sqrt{\psi_{\max}-\psi}}\right). \tag{4.13}$$

According to (2.25) the second factor in (4.13) is nonnegative. It remains to make sure that the third factor in (4.13) is nonpositive, i.e.,

$$\frac{f(t)}{4}+\frac{1}{\alpha}\frac{f'(t)}{t}\leq 0 \quad \forall t\in\left(0,\delta\sqrt{\psi_{\max}}\right). \tag{4.14}$$

To this end we introduce the function

$$F(t)=t^{4/\alpha}\left(\frac{f(t)}{4}+\frac{1}{\alpha}\frac{f'(t)}{t}\right). \tag{4.15}$$

Using (4.2) and (4.12) it is easily checked that

$$F'(t)=\frac{t^{4/\alpha}}{4}f'(t). \tag{4.16}$$

On the other hand we have from (4.2)

$$(t^a f')'=-t^a f\leq 0 \quad \forall t\in\left(0,\delta\sqrt{\psi_{\max}}\right). \tag{4.17}$$

The function $t^a f'(t)$ is therefore monotone decreasing in $(0,\delta\sqrt{\psi_{max}})$, which implies $f'(t)\leq 0$. In view of (4.16) we conclude then that

$$F(t)=\int_0^t F'(t)\,dt\leq 0, \tag{4.18}$$

and an insertion of this inequality into (4.13) completes the proof.

REFERENCES

[1] A. ACKER, L. E. PAYNE AND G. A. PHILIPPIN, *On the convexity of level lines of the fundamental mode in the fixed membrane problem, and the existence of convex solutions in a related free boundary problem*, Z. Angew. Math. Phys., 32 (1981), pp. 683–694.

[2] H. J. BRASCAMP AND E. H. LIEB, *Some inequalities for Gaussian measures and the long-range order of the one-dimensional plasma*, Functional Integration and Its Applications, A. Arthurs, ed., Oxford, 1975, pp. 1–14.

[3] L. S. FU AND L. WHEELER, *Stress bounds for bars in torsion*, J. Elasticity, 3 (1973), pp. 1–13.

[4] L. G. MAKAR-LIMANOV, *Solution of Dirichlet's problem for the equation $\Delta u=-1$ in a convex region*, Mathematical Notes of the Academy of Sciences of the USSR, 9 (1971), pp. 52–53.

[5] L. E. PAYNE, *Bounds for the maximal stress in the Saint-Venant torsion problem*, Ind. J. Mech. Math., special issue (1968), pp. 51–59.

[6] _____, *Bounds for solutions of a class of quasilinear elliptic boundary value problems in terms of the torsion function*, Proc. Royal Soc. Edinburgh, 88A (1981), pp. 251–265.

[7] L. E. PAYNE AND G. A. PHILIPPIN, *Some remarks on the problems of elastic torsion and of torsional creep*, Some Aspects of Mechanics of Continua, Part I, Jadavpur University, 1977, pp. 32–40.

[8] L. E. PAYNE AND I. STAKGOLD, *On the mean value of the fundamental mode in the fixed membrane problem*, Appl. Analysis, 3 (1973), pp. 295–303.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0008 $01.25/0

A BOUND FOR THE RATIO OF THE FIRST TWO EIGENVALUES OF A MEMBRANE*

GIUSEPPE CHITI†

Abstract. An isoperimetric inequality for the eigenfunctions of the Laplacian is proved, using rearrangements of functions. This result, together with a technique introduced by Payne, Polya and Weinberger (J. Math. Phys., 35 (1956), pp. 289–298), gives an upper bound (not isoperimetric) for the ratio of the first two eigenvalues of a membrane.

1. Introduction. Consider the equation of the vibrating membrane with fixed boundary,

$$\begin{aligned} &\Delta u+\lambda u=0 && \text{in } \Omega, \\ &u=0 && \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where Ω is a bounded domain in R^2. Denote by λ_n, $n=1,2,\cdots$ the sequence of eigenvalues of problem (1), numbered according to nondecreasing magnitude, and, by u_n, the sequence of the corresponding eigenfunctions. *Our aim is to find bounds for the ratio* λ_2/λ_1. This question is a particular aspect of a more general problem: to find inequalities among the first n eigenvalues of (1) which are independent of the geometry of the domain. The general problem is considered in [4] and [7], while in [1], [3], [6], the cases $n=2$, $n=3$ are examined. One of the results proved by Payne, Polya and Weinberger in [7], was the following:

$$\lambda_2\leq 3\lambda_1. \tag{2}$$

They also conjectured the inequality

$$\lambda_2\leq \alpha\lambda_1, \tag{3}$$

where α is the value for the disk and is approximately 2.538....In [1], Brands improved (2), obtaining

$$\lambda_2\leq 2.687\lambda_1, \tag{4}$$

and later de Vries [3] showed that

$$\lambda_2\leq 2.658\lambda_1. \tag{5}$$

We prove in this paper that

$$\lambda_2\leq 2.586\lambda_1. \tag{6}$$

The result is based essentially on the following

THEOREM. *Let u be an eigenfunction of problem* (1), *corresponding to the eigenvalue* λ. *Then u satisfies the inequality*

$$\int_\Omega u^2\,dx\leq\frac{3}{j_0^2-2}\lambda\int_\Omega |x|^2u^2\,dx, \tag{7}$$

*Received by the editors May 17, 1982, and in revised form September 20, 1982. This study was performed within the G.N.A.F.A. of the Italian Consiglio Nazionale delle Ricerche.

†Istituto Matematico Ulisse Dini, Viale Morgagni 67/A, 50134 Firenze, Italy.

where j_0 is the first positive zero of the Bessel function J_0. Equality holds in (7) *if and only if Ω is a disk (centered at the origin of R^2), and u is the first eigenfunction of* (1).

The inequality (7) can be easily extended to R^m, $m>2$; using the same methods we use here, one can see that (6) is the two-dimensional version of the following inequality:

$$\lambda_2 \le \left(1+\frac{m}{2}\,\frac{j_{(m/2)-1}^{-2} J_{m/2}^2\left(j_{(m/2)-1}\right)}{\int_0^1 r^3 J_{(m/2)-1}^2\left(j_{(m/2)-1} r\right) dr}\right)\lambda_1.$$

Also (7) can be extended to any power u^k of u, provided k is positive. This leads, in the case $m=2$, to the inequality

$$\lambda_2 \le \left(\frac{k^2}{2k-1}+\frac{2k}{2k-1}\,j_0^{-2}\,\frac{\int_0^1 r J_0^{2k}(j_0 r)\,dr}{\int_0^1 r^3 J_0^{2k}(j_0 r)\,dr}\right)\lambda_1,$$

valid for $k\ge 1$.

2. Rearrangements of functions. The technique we use to prove the theorem is a symmetrization technique, namely rearrangement of functions. Let u be a real measurable function defined in Ω, and consider the distribution function of u,

$$\mu(t)=\operatorname{meas}\{x\in\Omega:\ |u(x)|>t\}.$$

By definition, see [8], the *decreasing rearrangement* of u is the function

$$u^*(s)=\inf\{t\ge 0:\ \mu(t)<s\},$$

while the function $u^\star$,

$$u^\star(x)=u^*\left(\pi|x|^2\right)$$

is the *spherical rearrangement* of u. The domain of u^* is the interval $[0,M]$, where $M=\operatorname{meas}\Omega$; $u^\star$ is defined in the disk S, centered at the origin of R^2, such that $\operatorname{meas} S=M$. Since $u, u^*, u^\star$, all have the same distribution function, if u is in L^p, we have:

$$\int_\Omega u^p=\int_0^M u^{*p}=\int_S u^{\star p}. \tag{8}$$

3. From u to $u^\star$. In this section we study the behavior of the functional $\int_\Omega u^2/\int_\Omega |x|^2 u^2$ under symmetrization.

LEMMA 1. *Let u be a real measurable function defined in Ω. Then*

$$\frac{\int_\Omega u^2}{\int_\Omega |x|^2 u^2}\le\frac{\int_S u^{\star 2}}{\int_S |x|^2 u^{\star 2}}. \tag{9}$$

Proof. From (8) it follows that (9) is equivalent to

$$\int_S |x|^2 u^{\star 2}\le\int_\Omega |x|^2 u^2. \tag{10}$$

A well-known result of Hardy and Littlewood shows that

$$\int_\Omega |uv|\ge\int_0^M v_* u^* \tag{11}$$

for any pair of real measurable functions u and v; here v_* is the increasing rearrangement of v,

$$v_*(s)=v^*(M-s).$$

Replacing v with the function $|x|^2$, defined in Ω, we have:

$$\int_\Omega |x|^2 u^2 \geq \int_0^M |x|_*^2 u^{*2}. \tag{12}$$

From the definition of $|x|_*$, we obtain

$$|x|_*^2(s)\geq \frac{s}{\pi}, \qquad s\in[0,M],$$

and using (12) and the definition of u^*:

$$\int_\Omega |x|^2 u^2 \geq \int_0^M \frac{s}{\pi} u^{*2} = \int_S |x|^2 u^{*2}. \tag{13}$$

4. Comparison of u^* with a Bessel function. From now on we suppose that u is a solution of problem (1). The aim of this section is to compare u^* with the function

$$z(x)=J_0(\lambda^{1/2}|x|). \tag{14}$$

It is easy to verify that z is a solution of

$$\begin{aligned} &\Delta z+\lambda z=0 && \text{in } S_\lambda,\\ &z=0 && \text{on } \partial S_\lambda, \end{aligned} \tag{15}$$

where

$$S_\lambda=\{x\in R^2: |x|\leq j_0/\lambda^{1/2}\}. \tag{16}$$

LEMMA 2. *Let u be a solution of problem* (1), *corresponding to the eigenvalue* λ, *and let z be the function defined in* (14). *Then*:

$$\frac{\int_S u^{*2}}{\int_S |x|^2 u^{*2}} \leq \frac{\int_{S_\lambda} z^2}{\int_{S_\lambda} |x|^2 z^2}. \tag{17}$$

Proof. We normalize u in such a way that

$$\int_S u^{*2} = \int_{S_\lambda} z^2. \tag{18}$$

In [2, §4] the following result is proved. If condition (18) holds and $M_\lambda=\operatorname{meas} S_\lambda$, $M=\operatorname{meas}\Omega$, then: a) $M_\lambda\leq M$; b) there exists s_1, $s_1\in(0,M_\lambda)$, such that

$$\begin{aligned} &z^*(s)\geq u^*(s), && s\in[0,s_1],\\ &z^*(s)\leq u^*(s), && s\in[s_1,M_\lambda]. \end{aligned} \tag{19}$$

From (18) and (19) we have

$$\int_0^{M_\lambda} sz^{*2} - \int_0^{M} su^{*2} = \int_0^{s_1} s(z^{*2}-u^{*2}) + \int_{s_1}^{M_\lambda} s(z^{*2}-u^{*2}) - \int_{M_\lambda}^{M} su^{*2}$$

$$\leq s_1 \int_0^{s_1} (z^{*2}-u^{*2}) + s_1 \int_{s_1}^{M_\lambda} (z^{*2}-u^{*2}) - s_1 \int_{M_\lambda}^{M} u^{*2}$$

$$= s_1 \left[\int_0^{M_\lambda} z^{*2} - \int_0^{M} u^{*2} \right] = 0.$$

This, together with (18), concludes the proof, since

$$\int_S |x|^2 u^{*2} = \frac{1}{\pi} \int_0^M su^{*2},$$

and

$$\int_{S_\lambda} |x|^2 z^2 = \frac{1}{\pi} \int_0^M sz^{*2}.$$

5. Proof of the theorem. Using Lemma 1 and Lemma 2, and the definition of z, we have

$$\frac{\int_\Omega u^2}{\int_\Omega |x|^2 u^2} \leq j_0^{-2} \frac{\int_0^1 r J_0^2(j_0 r)\, dr}{\int_0^1 r^3 J_0^2(j_0 r)\, dr} \lambda. \tag{20}$$

The value of the right-hand side can be computed, using for instance, the formulas in [5, p. 262], and we obtain (7).

6. A bound for λ_2/λ_1. This section is devoted to an application of the theorem, which enables us to obtain an upper bound for the ratio λ_2/λ_1 of the first two eigenvalues of problem (1). Let $u_1(x_1, x_2)$ be the first eigenfunction of (1). We choose coordinate axes such that

$$\int_\Omega x_1 u_1^2\, dx = \int_\Omega x_2 u_1^2\, dx = 0.$$

With this choice, since the functions $x_1 u_1$ and $x_2 u_1$ are orthogonal to u_1, we have

$$\lambda_2 \leq \frac{\int_\Omega |\operatorname{grad} x_i u_1|^2\, dx}{\int_\Omega x_i^2 u_1^2\, dx}, \qquad i = 1, 2, \tag{21}$$

which implies

$$\lambda_2 \leq \frac{\int_\Omega \left(|\operatorname{grad} x_1 u_1|^2 + |\operatorname{grad} x_2 u_1|^2 \right) dx}{\int_\Omega |x|^2 u_1^2\, dx}. \tag{22}$$

Integrating twice by parts, using the fact that u_1 solves problem (1), we obtain

$$\int_\Omega \left(|\operatorname{grad} x_1 u_1|^2 + |\operatorname{grad} x_2 u_1|^2 \right) dx = \int_\Omega |x|^2 |\operatorname{grad} u_1|^2 dx = \lambda_1 \int_\Omega |x|^2 u_1^2\, dx + 2 \int_\Omega u_1^2\, dx,$$

and from (22)

$$\lambda_2 \leq \lambda_1 + 2\frac{\int_\Omega u_1^2\,dx}{\int_\Omega |x|^2 u_1^2\,dx}.$$

From (7) we have

$$\lambda_2 \leq \lambda_1\left(1+\frac{6}{j_0^2-2}\right);$$

since $j_0 > 2.40482$, this implies (6).

Acknowledgment. The author wishes to thank Prof. Murray H. Protter, who introduced him to this subject; thanks are also due to Dr. M. G. Gasparo, who helped handle Bessel functions.

REFERENCES

[1] J. J. A. M. BRANDS, *Bounds for the ratios of the first three membrane eigenvalues*, Arch. Rational Mech. Anal., 16 (1964), pp. 265–268.

[2] G. CHITI, *A reverse Hölder inequality for the eigenfunctions of linear second order elliptic operators*, Z. Angew. Math. Phys., 33 (1982), pp. 143–148.

[3] H. L. DE VRIES, *On the upper bound for the ratio of the first two membrane eigenvalues*, Z. Naturforsch, 22 (1967), pp. 152–153.

[4] G. N. HILE AND M. H. PROTTER, *Inequalities for eigenvalues of the Laplacian*, Indiana Univ. Math. J., 29 (1980), pp. 523–538.

[5] Y. L. LUKE, *Integrals of Bessel Functions*, McGraw-Hill, New York, 1962.

[6] P. MARCELLINI, *Bounds for the third membrane eigenvalue*, J. Differential Equations, 37 (1980), pp. 438–443.

[7] L. E. PAYNE, G. POLYA AND H. F. WEINBERGER, *On the ratio of consecutive eigenvalues*, J. Math. Phys., 35 (1956), pp. 289–298.

[8] G. TALENTI, *Elliptic equations and rearrangements*, Ann. Scuola Norm. Sup. Pisa, (3) 4 (1976), pp. 697–718.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0009 $01.25/0

SINGULAR PERTURBATIONS FOR A SEMILINEAR HYPERBOLIC EQUATION*

GEORGE C. HSIAO† AND RICHARD J. WEINACHT†

Abstract. The Cauchy problem for a semilinear hyperbolic equation with a small parameter is considered. The reduced problem is of parabolic type and, although there is no reduction of order, there is an initial layer. An asymptotic solution with boundary layer corrections is constructed and, for a restricted class of nonlinearities, is shown to be uniformly asymptotically valid for sets bounded in the time direction.

1. Introduction. We consider the pure initial value problem for the semilinear hyperbolic equation

$$L_\varepsilon[u] := \varepsilon^2 u_{tt} + u_t - u_{xx} = F(u), \qquad t>0, \tag{1.1}$$

where ε is a small positive parameter. The Cauchy data are

$$u(x,0;\varepsilon) = f(x), \tag{1.2}$$

$$\varepsilon u_t(x,0;\varepsilon) = g(x), \qquad x\in\mathbb{R}. \tag{1.3}$$

For $\varepsilon=0$, (1.1) becomes the semilinear parabolic equation

$$U_t - U_{xx} = F(U), \qquad t>0, \tag{1.4}$$

for which the single initial condition

$$U(x,0) = f(x) \tag{1.5}$$

is appropriate for a well-posed problem.

Thus there is a boundary layer at $t=0$. By means of a composite expansion method [19] we construct an asymptotic solution to all orders in ε. For a restricted class of nonlinearities F and initial data f, g, we prove uniform asymptotic validity on regions bounded in the t direction. Our main result embodying this assertion is formulated in the theorem in §5.

For linear problems (e.g. $F\equiv 0$) "hyperbolic-parabolic" perturbations of the type treated here have been considered by several authors, including Zlamal [20], [21], Bobisud [2] and Bensoussan–Lions–Papanicolaou [1]. In particular the present authors treated the problem (1.1)–(1.3) for $F\equiv 0$ in [9], and for weakly nonlinear $F=\varepsilon^2\hat{F}$ in [10]. A physical interpretation of the problem (1.1)–(1.3) in a purely mechanical setting is given in [9],[10]. In addition, for $F\equiv 0$, (1.1) is precisely the "wave equation of heat conduction" proposed by several authors (see Nowinski [16, Chap. 7] and the references therein) to oversome the shortcoming of the usual equation of heat conduction ($\varepsilon=0$), which predicts infinite speeds of propagation. Our results confirm in a rigorous (and more quantitative) way previous conclusions that the perturbation term $\varepsilon^2 u_{tt}$ has an appreciable effect only for small times. Moreover, "initial layer corrections" are given, to all orders.

Singular perturbation problems for hyperbolic equations where there is a loss of order in the reduced equation have been treated by de Jager [12], Geel and de Jager [7]

* Received by the editors October 16, 1981, and in revised form September 23, 1982.
† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711.

and Geel [6] as well as Genet and Maudaune [8]. As in the work of de Jager and Geel, our proofs depend upon energy estimates and a fixed point theorem. There is a considerable difference in the details, however, since the characteristics in our case depend on ε.

The overall restrictions on the nonlinearity F are the monotonicity condition

$$F'(z)<0 \quad \forall z\in\mathbb{R} \tag{1.6}$$

and, from a physical point of view, the very reasonable condition

$$F(0)=0, \tag{1.7}$$

as well as smoothness up to an order depending on the order of the approximation (see the theorem in §5 for a precise statement). Our main result gives uniform asymptotic validity on any strip $\mathbb{R}\times[0,t_0]$ for which the reduced problem (1.4)–(1.5) has a bounded classical solution. For smooth F, of course, there is always a local (i.e. for small time) solution of the reduced problem. If, in addition, the boundedness of F' on $\mathbb{R}$ is assumed, then our results become global, i.e. on any such strip. Naturally, this limits the growth of F at ∞ (even without invoking (1.7)):

$$|F(z)|\leq\lambda|z|+\mu.$$

Similar growth limitations are considered by Brézis–Nirenberg [3] (see also Cesari–Kannan [4]), who are not concerned with singular perturbations.

Results of this paper have been extended recently by Esham [5] to nonlinear evolution equations in Hilbert space.

A previous version of the present work, but with restrictions on the growth of the nonlinearity at infinity, was presented at the Conference on Nonlinear Partial Differential Equations in Engineering and Applied Science at the University of Rhode Island in June, 1979. The results of the present version were reported at the Annual Meeting of the American Mathematical Society in San Francisco in January, 1981 (Abstract 783-35-32).

2. Formal expansion. Guided by our results [9] for the case $F\equiv 0$, we use the stretched variable $\tilde{t}=t/\varepsilon^2$ and make the following ansatz for the solution u of (1.1)–(1.3):

$$u(x,t;\varepsilon)\sim\sum_{n=0}^{\infty}\varepsilon^n\left[U_n(x,t)+\varepsilon V_n(x,\tilde{t})\right],$$

which yields upon substitution into (1.1)–(1.3) the parabolic initial value problems for the U_n:

$$U_{n,t}-U_{n,xx}=\begin{cases}F(U_0), & n=0,\\ F'(U_0)U_1, & n=1,\\ F_n-U_{n-2,tt}, & n\geq 2\end{cases}\quad (x,t)\in\mathbb{R}\times(0,\infty), \tag{2.1}$$

$$U_n(x,0)=\begin{cases}f(x), & n=0,\\ -V_{n-1}(x,0), & n\geq 1,\end{cases}\quad x\in\mathbb{R}, \tag{2.2}$$

and the ODE problems for the V_n:

$$V_{n,\tilde{t}\tilde{t}}+V_{n,\tilde{t}}=\begin{cases}0, & n=0,1,\\ G_n+V_{n-2,xx}, & n\geq 2,\end{cases}\qquad \tilde{t}>0, \tag{2.3}$$

$$V_{n,\tilde{t}}(x,0)=\begin{cases}g(x), & n=0,\\ -U_{n-1,t}(x,0), & n\geq 1,\end{cases}\qquad x\in\mathbb{R}, \tag{2.4}$$

together with the matching condition

$$V_n(x,\tilde{t})\to 0 \quad \text{as } \tilde{t}\to\infty \tag{2.5}$$

for all $n\geq 0$. The terms F_n and G_n in multi-index notation are

$$F_n:=\sum \frac{F^{|\alpha|}(U_0(x,t))}{|\alpha|!}\binom{|\alpha|}{\alpha}[U_1(x,t)]^{\alpha_1}\cdots[U_n(x,t)]^{\alpha_n},$$

where the sum is over all α such that

$$\sum_{i=1}^{n} i\alpha_i=n$$

and

$$G_n:=\sum \frac{\binom{|\beta|}{\beta}\binom{|\gamma|}{\gamma}}{|\beta|!|\gamma|!}F^{(|\beta|+|\gamma|)}\big(U_0(x,0)[V_0(x,\tilde{t})]^{\beta_1}\cdots [V_{n-2}(x,\tilde{t})]^{\beta_{n-1}}[W_1(x,\tilde{t})]^{\gamma_1}\cdots[W_{n-1}(x,\tilde{t})]^{\gamma_{n-1}}\big),$$

where the last sum is over all β,γ with $\beta\neq 0$ such that

$$\sum_{i=1}^{n-1} i(\beta_i+\gamma_i)=n-1$$

and

$$W_k(x,\tilde{t})=\sum_{m=0}^{[k/2]} D_t^m U_{k-2m}(x,0)\tilde{t}^m/m!.$$

Thus all the problems for the U_n, V_n are linear except that for U_0 which satisfies a semilinear IVP for the heat operator with smooth initial data f. Partially inverting the problem for U_0 leads to consideration of the nonlinear integral equation

$$v=K[f]+K_1[F(v)], \tag{2.6}$$

where K and K_1 are the familiar Poisson

$$K[\omega](x,t):=\int_{-\infty}^{\infty}\gamma(x-\xi,t)\omega(\xi)\,d\xi$$

and the Duhamel–Poisson operators

$$K_1[\omega](x,t) := \int_0^t \int_{-\infty}^{\infty} \gamma(x-\xi, t-\tau)\omega(\xi,\tau)\, d\xi\, d\tau$$

respectively, and

$$\gamma(z,t) := (4\pi t)^{-1/2} \exp\{-z^2/4t\}$$

is the fundamental solution of the heat operator in one space dimension.

If f and F' are continuous on $\mathbb{R}$, and f is bounded on $\mathbb{R}$, then, via the contraction mapping principle, it is easy to establish locally (i.e. on some strip $S_0 := \mathbb{R} \times [0, t_0]$, $t_0 > 0$) the existence and uniqueness of a continuous bounded solution of (2.6). Such a solution is necessarily a classical solution of (2.1)–(2.2) for $n=0$ on S_0. Moreover, from (2.6), one can bound the derivatives of U_0 on S_0 (see the lemma below), provided the data f and F are sufficiently smooth. If, in addition, F' is bounded on $\mathbb{R}$, then the results are global, i.e. on $R \times [0, \infty)$.

If such a U_0 is known on *any* such strip S_0 (perhaps larger than guaranteed by the above argument), the problems for the remaining U_n ($n \geq 1$) and all the V_n are linear and their solutions are global on S_0, and one obtains representations and estimates for the U_n, V_n and their derivatives analogous to those [9] for the linear case. It is to be emphasized that in the problems (2.1)–(2.2) and (2.3)–(2.5), the derivatives of F are evaluated only at $U_0(x,t)$ as (x,t) ranges over S_0. Hence the continuity of $F^{(j)}$ on $\mathbb{R}$ ensures the boundedness of $F^{(j)}(U_0(x,t))$ on S_0.

We summarize these results in

LEMMA 1. *Assume for $n=0$ that* (2.1)–(2.2) *has a unique bounded classical solution U_0 on $S_0 := \mathbb{R} \times [0, t_0]$. Let l, m and n be nonnegative integers and assume that f, g belong to $C^s(\mathbb{R})$ with $f^{(j)}, g^{(j)}$ bounded on $\mathbb{R}$ for $0 \leq j \leq s$, $s = 2n+l+2m$. Suppose F belongs to $C^r(\mathbb{R})$ with $r = n+l+2m$. Then there exists a constant $\mathcal{C}$, independent of x, t (and ε), and a polynomial P dependent on l, m, n, such that on S_0*

$$|D_x^l D_t^m U_n(x,t)| \leq \mathcal{C}\big(\|\phi^{(s)}\|, \|F^{(r)}\|; t_0\big), \tag{2.7}$$

$$|D_x^l D_t^m V_n(x,\tilde{t})| \leq e^{-\tilde{t}} P(\tilde{t}) \mathcal{C}\big(\|\phi^{(2n+l)}\|, \|F^{(n+l-1)}\|; t_0\big). \tag{2.8}$$

Moreover for $t=0$

$$|D_x^l D_t^m U_n(x,0)| \leq \mathcal{C}\big(\|\phi^{(s)}\|, \|F^{r-2}\|; t_0\big). \tag{2.9}$$

Here $\|\phi^{(j)}\|$ denotes the supremum norm over $\mathbb{R}$ of all derivatives (including the function itself) of f, g up through order j and $\|F^{(j)}\|$ indicates the maximum over $1 \leq l \leq j$ of

$$\sup |F^{(l)}(U_0(x,t))|$$

as (x,t) ranges over S_0. If $j \leq 0$ (as happens when $l=m=n=0$), $\|F^{(j)}\|$ is to be replaced by unity.

The long induction proof of these results uses in a key way (2.6) and Gronwall's inequality. The induction step for V_{n+1} is based upon the estimate

$$|D_{\tilde{t}}^m D_x^l V_{n+1}(x,t)| \leq e^{-\tilde{t}} \cdot \left\{ |D_x^l U_{n,t}(x,0)| + Q(\tilde{t}) \sum_{j=0}^{m-2} \left(|D_{\tilde{t}}^j D_x^l \hat{G}_{n+1}| + |D_{\tilde{t}}^j D_x^{l+2} \hat{V}_{n-1}| \right) \right\},$$

where Q is a polynomial, $\hat{V}_{n-1}=e^{t}\tilde{V}_{n-1}$ and $\hat{G}_{n+1}=e^{t}\tilde{G}_{n+1}$. The induction step for U_{n+1} is based upon the estimate

$$\begin{aligned}|D_t^m D_x^l U_{n+1}(x,t)|\leq \sup_{S_0}\Big\{&|D_x^{l+m}U_{n+1}(x,0)|\\ &+\sum_{j=1}^{m}\Big(|D_t^{m-j}D_x^{2j-2+l}\hat{F}_{n+1}|+|D_t^{m-j+2}D_x^{2j-2+l}U_{n-1}|\\ &\qquad+|D_t^{m-j}D_x^{2j-2+l}(F'(U_0)U_{n+1})|\Big)\\ &+t\Big(|D_x^{2m+l}\hat{F}_{n+1}|+|D_x^{2m+l}U_{m-1,tt}|+|D_x^{2m+l}(F'(U_0)U_{n+1})|\Big)\Big\},\end{aligned}$$

where $\hat{F}_{n+1}$ is the F_{n+1} appearing in (2.1) with the highest order term $F'(U_0)U_{n+1}$ missing. In carrying out the induction proof we find it necessary to hypothesize (2.9) separately.

Remark. If one assumes also that F' is bounded on $\mathbb{R}$, then the above results are global, i.e. on S_0 for any t_0.

3. A priori estimates. In this section we use energy integrals to obtain the a priori estimates (3.2) and (3.3) related to the solution of the initial value problem

$$\begin{aligned}\tilde{L}_\varepsilon[Z]:=L_\varepsilon[Z]+q(x,t;\varepsilon)Z=h(x,t;\varepsilon),\qquad t>0,\\ Z(x,0;\varepsilon)=0,\\ \varepsilon Z_t(x,0;\varepsilon)=\psi(x;\varepsilon),\end{aligned}$$

where q is a given positive continuous function which is bounded and bounded away from zero,

$$0<q_0\leq q(x,t;\varepsilon)\leq q_1<\infty, \tag{3.1}$$

for small ε and for all real x, and $0\leq t\leq t_0$ with q_0, q_1 independent of ε. The method of proof is essentially the *abc*-method of Friedrichs (Protter [17], Morawetz [15]), but with adaptations necessary due to the dependence of the characteristics of L_ε on ε. The use of energy integrals for estimates in hyperbolic singular perturbation problems seems first to have appeared in de Jager [12] (see also Geel and de Jager [7], Geel [6]), where there is a loss of order in the reduced equation, and the characteristics are independent of ε.

We state our result as

LEMMA 2. *Let Q denote the open rectangular region $(\alpha,\beta)\times(0,t_0)$. Then there exist positive constants ε_0, b depending only on q_0, q_1 (given in* (3.1)*), such that for any $C^2(\bar{Q})$ function Z which vanishes on both the bottom $t=0$ and also on the sides $x=\alpha$ and $x=\beta$ of Q, one has*

$$|Z(x,t;\varepsilon)|^2\leq[\lambda(\varepsilon)]^{-1}\left\{\frac{b}{2}\int_\alpha^\beta|\varepsilon Z_t(\xi,0)|^2d\xi+\int_0^{t_0}\int_\alpha^\beta|\tilde{L}_\varepsilon[Z]|^2d\xi d\tau\right\} \tag{3.2}$$

for all (x,t) in Q, $0<\varepsilon\leq\varepsilon_0$, and where

$$\lambda(\varepsilon)=\tfrac{1}{2}b\varepsilon^2+O(\varepsilon^4).$$

Moreover, the usual $H^1(Q)$ *norm of* Z,

$$\|Z\|^2_{H^1(Q)} := \int_0^{t_0}\int_\alpha^\beta \left(|Z|^2+|Z_t|^2+|Z_x|^2\right) d\xi\, d\tau,$$

satisfies

$$\|Z\|^2_{H^1(Q)} \le t_0 \cdot \text{RHS}, \tag{3.3}$$

where RHS *is the right-hand side of* (3.2).

Remark. Of course, by density, the inequalities (3.2), (3.3) hold for a broader class of functions than stated in Lemma 2.

Proof. For brevity, define

$$h(x,t;\varepsilon) := \tilde{L}_\varepsilon[Z](x,t),$$
$$\psi(x;\varepsilon) := \varepsilon Z_t(x,0).$$

Then for arbitrary constants a, b (independent of x,t) to be specified later, integration by parts of the identity for $0<t<t_0$,

$$0=\int_0^t\int_\alpha^\beta (aZ+bZ_t)\left(L_\varepsilon[Z]+qZ-h\right) d\xi\, d\tau,$$

yields our basic identity

$$\int_\alpha^\beta (\mathbf{Z}'A\mathbf{Z})(\xi,t)\, d\xi = \frac{b}{2}\int_\alpha^\beta |\psi(\xi;\varepsilon)|^2 d\xi + \int_0^t\int_\alpha^\beta (h\Gamma'\mathbf{Z}-\mathbf{Z}'B\mathbf{Z})\, d\xi\, d\tau,$$

where we have introduced the vectors $\mathbf{Z} := \operatorname{col}(Z, Z_t, Z_x)$, $\Gamma := \operatorname{col}(a,b,0)$ and the symmetric matrices

$$A := \frac{1}{2}\begin{pmatrix} a & a\varepsilon^2 & 0 \\ a\varepsilon^2 & b\varepsilon^2 & 0 \\ 0 & 0 & b \end{pmatrix}, \qquad B := \begin{pmatrix} aq & bq/2 & 0 \\ bq/2 & b-a\varepsilon^2 & 0 \\ 0 & 0 & a \end{pmatrix},$$

and prime denotes transpose of a vector.

The matrix A is positive definite if and only if $b>a\varepsilon^2>0$. Under these conditions the lowest (necessarily positive) eigenvalue $\lambda(\varepsilon)$ of A is given for small ε by

$$\lambda(\varepsilon) = (b/2)\varepsilon^2 + O(\varepsilon^4)$$

as an elementary computation shows. Then by use of the arithmetic-geometric mean inequality we arrive at our basic inequality

$$\lambda(\varepsilon)\int_\alpha^\beta (\mathbf{Z}'\mathbf{Z})(\xi,t)\, d\xi \le \int_\alpha^\beta (\mathbf{Z}'A\mathbf{Z})(\xi,t)\, d\xi \tag{3.4}$$
$$\le \frac{b}{2}\int_\alpha^\beta |\psi(\xi;\varepsilon)|^2 d\xi + \int_0^t\int_\alpha^\beta \left(|h|^2-\mathbf{Z}'C\mathbf{Z}\right) d\xi\, d\tau,$$

where the symmetric matrix C is defined by

$$C := \begin{pmatrix} aq-a^2/2 & bq/2 & 0 \\ bq/2 & b-a\varepsilon^2-b^2/2 & 0 \\ 0 & 0 & a \end{pmatrix}.$$

If a, b can be chosen according to the above restriction and such that C is positive semidefinite, then (3.4) becomes

$$\lambda(\varepsilon)\int_\alpha^\beta (\mathbf{Z}'\mathbf{Z})(\xi,t)\,d\xi \le \frac{b}{2}\int_\alpha^\beta |\psi(\xi;\varepsilon)|^2 d\xi + \int_0^{t_0}\int_\alpha^\beta |h|^2 d\xi\, d\tau.$$

Hence by integration of both sides of this inequality with respect to t from 0 to t_0, one obtains (3.3). On the other hand, by use of the elementary inequality

$$|Z(x,t)|^2 \le \int_\alpha^x \left(|Z|^2 + |Z_x|^2\right) d\xi,$$

we arrive at the pointwise estimate (3.2). In each case b is independent of ε on $(0, \varepsilon_0]$), with ε_0 determined as follows.

The matrix C is pointwise semi-definite for given a, b with $b > a\varepsilon^2 > 0$, if and only if

$$q \ge a/2, \qquad a(q-a/2)(b-a\varepsilon^2-b^2/2) \ge b^2q^2/4. \tag{3.5}$$

An elementary computation shows that for $a = q_0$ (so that the first inequality in (3.5) is satisfied), the second inequality in (3.5) is satisfied, provided that b is chosen between the two positive roots of the quadratic function

$$J(b) := \left[q^2 + q_0(2q-q_0)\right]b^2 - 2q_0(2q-q_0)b + 2q_0^2(2q-q_0)\varepsilon^2.$$

A careful but elementary analysis of the roots shows that there exists a positive ε_0 so that for $0 < \varepsilon \le \varepsilon_0$ the quantity b can be chosen independent of ε on $(0, \varepsilon_0]$, but depending on q_0, q_1 such that $J(b) \le 0$. The method fails if $q_0 = 0$ or $q_1 \to +\infty$. This completes the proof of the lemma.

4. A related linear problem. To justify our formal asymptotic result we consider here the linear initial-boundary value problem

$$\tilde{L}_\varepsilon[v] \equiv L_\varepsilon[v] + q(x,t;\varepsilon)v = h(x,t;\varepsilon) \quad \text{in } Q, \tag{4.1}$$

$$v(x,0) = v_t(x,0) = 0, \qquad \alpha \le x \le \beta, \tag{4.2}$$

$$v(\alpha,t) = v(\beta,t) = 0, \qquad 0 \le t \le t_0, \tag{4.3}$$

where Q is the open rectangular region $(\alpha,\beta)\times(0,t_0)$.

The a priori estimate of §3 is closely related to this problem and is used in the convergence proof of our Galerkin approximations.

For problem (4.1)–(4.3) we use the existence/uniqueness and regularity results embodied in Lemma 3, below. For $s \ge 1$ let $\tilde{H}^s(Q)$ denote the subspace of functions which belong to the usual Sobolev space $H^s(Q)$ and which vanish on the bottom and vertical sides of Q. Then letting ε_0 denote the positive number in Lemma 2 of §3, our result is

LEMMA 3. *Suppose for each ε on $(0, \varepsilon_0]$, the function $q \in C(\bar{Q})$. Then for each $h \in L_2(Q)$ there exists a unique generalized solution in $\tilde{H}^1(Q)$ of the* IBVP (4.1)–(4.3). *Moreover for sufficiently smooth q and h (say $q, h \in C^3(\bar{Q})$) with h vanishing in a boundary strip of the vertical sides of Q, the generalized solution of* (4.1)–(4.3) *is classical.*

For q independent of t one can use the Fourier method to establish the assertions of Lemma 3 (see e.g. [11]). For q depending on t, however, an explicit reference for the proof of Lemma 3 seems not to be available. Therefore, we have applied the Galerkin method to prove the lemma. The existence/uniqueness of a generalized solution of a

similar problem but with q independent of t appears in Mikhailov [14, pp. 299–305]. The modifications to the present case are clear from the Galerkin scheme set forth there. A boot-strap argument based on the Galerkin equations then yields the regularity result. For the latter result the vanishing of h near the lower corners of Q is vital.

From Lemma 3 it follows that the IBVP (4.1)–(4.3) defines an (inverse) mapping Λ_ε from $L_2(Q)$ into $\tilde{H}^1(Q)$. Moreover, again from Lemma 3, for smooth q the restriction of Λ_ε to $C^3(\bar{Q})$ functions which vanish identically in a boundary strip of the vertical sides maps such functions into classical solutions of (4.1)–(4.3).

5. Main result. With the above preparations we can now state our main result in

THEOREM 1. *Assume*

$$\text{(i)} \qquad F'(z)<0 \quad \forall z\in\mathbb{R},$$

$$\text{(ii)} \qquad F\in C^r(\mathbb{R}), \qquad r=\max(8,N+6),$$

$$\text{(iii)} \qquad F(0)=0,$$

$$\text{(iv)} \qquad f,g\in H^s(\mathbb{R}), \qquad s=2N+9.$$

In addition, suppose that the reduced problem (1.4)–(1.5) *has a bounded classical solution on the strip* $S_0=\mathbb{R}\times[0,t_0]$, $t_0>0$. *Then the solution* u *of* (1.1)–(1.3) *admits the asymptotic representation*

$$u(x,t;\varepsilon)=U_0(x,t)+\sum_{n=1}^{N}\varepsilon^n\left[U_n(x,t)+V_{n-1}(x,t/\varepsilon^2)\right]+O(\varepsilon^{N+1}) \tag{5.1}$$

as $\varepsilon\to 0^+$ *uniformly on* S_0 *where* N *is any nonnegative integer.*

Remark. The restrictions (i)–(iv) on F,f and g will ensure that (1.4)–(1.5) and (1.1)–(1.3) have local solutions, i.e. on some strip $\mathbb{R}\times[0,t_0]$. With the additional assumption that (1.4)–(1.5) has a bounded classical solution on any strip, our proof will show that (1.1)–(1.3) has a classical solution on the same strip for sufficiently small ε.

Proof. The proof consists of examining the remainder $Z:=u-[U+\varepsilon V]$, where, for ν to be chosen,

$$U(x,t;\varepsilon):=\sum_{l=0}^{\nu}\varepsilon^l U_l(x,t),$$

$$V(x,t;\varepsilon):=\sum_{l=0}^{\nu-1}\varepsilon^l V_l(x,t/\varepsilon^2)$$

and the U_l, V_l are the solutions of (2.1)–(2.2) and (2.3)–(2.5) respectively. Here $0<\varepsilon\le\varepsilon_0$, with ε_0 given in §3. In order to deal with homogeneous initial conditions, put $w:=Z-X$ where

$$X(x,t;\varepsilon):=(t/\varepsilon)\psi(x;\varepsilon)e^{-t/\varepsilon^2}$$

with

$$\psi(x;\varepsilon):=-\varepsilon^\nu U_{\nu-1,t}(x,0)-\varepsilon^{\nu+1}U_{\nu,t}(x,0).$$

Then w is a classical solution of the initial value problem for

$$L_\varepsilon[w]-F'(U+\varepsilon V+X)w=G(w)+h \tag{5.2}$$

with homogeneous initial conditions $w(x,0)=w_t(x,0)=0$, and the known nonhomogeneous term h is given by

$$h := [F(U+\varepsilon V+X)-F(U+\varepsilon V)] + [F(U+\varepsilon V)-F(U)-\varepsilon L_\varepsilon[V]]+[F(U)-L_\varepsilon[U]]-L_\varepsilon[X], \tag{5.3}$$

while the nonlinear term G is

$$G(w) := F(U+\varepsilon V+X+w)-F(U+\varepsilon V+X)-F'(U+\varepsilon V+X)w, \tag{5.4}$$

i.e. the difference between F at u and a linear approximation thereto.

We consider now a related nonlinear initial-boundary value problem (IBVP). For any fixed positive ε and (x_0,t_0) with $t_0>0$, introduce a $C^\infty(\mathbb{R})$ cut-off function $\zeta=\zeta(x)$, which is identically one on the interval $[x_0-t_0/\varepsilon, x_0+t_0/\varepsilon]$ on the initial line cut off by the characteristics of L_ε extending backward through (x_0,t_0), and which vanishes identically on $(-\infty, x_0-1-t_0/\varepsilon]$ and on $[x_0+1+t_0/\varepsilon,\infty)$. Now let $\alpha=\alpha(\varepsilon)<x_0-1-2t_0/\varepsilon$ and $\beta=\beta(\varepsilon)>x_0+1+2t_0/\varepsilon$, and consider the open rectangular region $Q:=(\alpha,\beta)\times(0,t_0)$. The IBVP in question is

$$\tilde{L}_\varepsilon[\bar{w}]=\bar{G}(\bar{w})+\bar{h}, \quad \text{in } Q, \tag{5.5}$$

$$\bar{w}(x,0)=\bar{w}_t(x,0)=0, \qquad \alpha\le x\le\beta, \tag{5.6}$$

$$\bar{w}(\alpha,t)=\bar{w}(\beta,t)=0, \qquad 0\le t\le t_0, \tag{5.7}$$

where $\bar{G}=\zeta G$ and $\bar{h}=\zeta h$. The operator $\tilde{L}_\varepsilon$ in (5.5) is the same as that in (4.1) (also in §3) with

$$q(x,t;\varepsilon)\equiv -F'(U(x,t;\varepsilon)+\varepsilon V(x,t;\varepsilon)+X(x,t;\varepsilon)).$$

Since F' is continuous and negative on $\mathbb{R}$, then the q given here indeed satisfies the condition (3.1) for $\varepsilon\in(0,\varepsilon_0]$, due to the boundedness of $U+\varepsilon V+X$ on the strip $\mathbb{R}\times[0,t_0]$. Therefore the estimates of §3 apply to $\tilde{L}_\varepsilon$ so that the inverse operator Λ_ε of §4 may be used. Hence we formally convert the IBVP (5.5)–(5.7) to the operator equation

$$\bar{w}=\Lambda_\varepsilon\bar{G}(\bar{w})+\Lambda_\varepsilon\bar{h},$$

and we are led to consider fixed points of the operator A_ε where

$$A_\varepsilon(v) := \Lambda_\varepsilon\bar{G}(v)+\Lambda_\varepsilon\bar{h}. \tag{5.8}$$

In the appendix it is shown that A_ε has a unique fixed point w^* which turns out to be a classical solution of the IBVP (5.5)–(5.7). Moreover, in the characteristic triangle T_ε, the triangular region bounded below by the initial line $t=0$ and above by the two characteristics of L_ε extending backward through (x_0,t_0), the solution w of the IVP for (5.2) with homogeneous initial conditions and w^* coincide.

The pointwise estimate (3.2) applied to w^* yields

$$|w^*(x,t;\varepsilon)|^2\le[\lambda(\varepsilon)]^{-1}\int_0^t\int_\alpha^\beta|\bar{G}(w^*)+\bar{h}|^2\,d\xi\,d\tau,$$

so that from (A.1) in the Appendix and the arithmetic-geometric mean inequality

$$\sup_{\alpha\le x\le\beta}|w^*(x,t;\varepsilon)|^2\le\frac{1}{2}[\lambda(\varepsilon)]^{-1}(\beta-\alpha)\|F''\|_*^2\int_0^t\sup_{\alpha\le x\le\beta}|w^*(x,\tau;\varepsilon)|^4d\tau$$

$$+2[\lambda(\varepsilon)]^{-1}\|\bar{h}\|_2^2,$$

and hence from a quadratic Gronwall inequality (see e.g. [18]), using $\|\cdot\|_p$ to denote the $L_p(Q)$ norm,

$$\sup_{\alpha\le x\le\beta}|w^*(x,t;\varepsilon)|^2\le\frac{2[\lambda(\varepsilon)]^{-1}\|\bar{h}\|_2^2}{1-\gamma t_0}, \tag{5.9}$$

where $\gamma:=(\beta-\alpha)[\lambda(\varepsilon)]^{-2}\|F''\|_*^2\|\bar{h}\|_2^2$.

Taking into account the problems satisfied by U_l, V_l, one sees from (5.3) that

$$|\bar{h}|\le e^{-t/\varepsilon^2}\Big\{\|F'\|_*(t/\varepsilon)|\psi(x,\varepsilon)|+\varepsilon\big[|\psi(x;\varepsilon)|+|\psi''(x,\varepsilon)|t/\varepsilon^2\big]\Big\}$$

$$+\left|\varepsilon^{\nu-1}V_{\nu-2,xx}+\varepsilon^{\nu}V_{\nu-1,xx}-\varepsilon^{\nu+1}U_{\nu-1,tt}-\varepsilon^{\nu+2}U_{\nu,tt}\right|.$$

At this point the estimates of the L_2 norm of $\bar{h}$ proceed as in the linear case [9, p. 249] so that $\|\bar{h}\|_2^2=O(\varepsilon^{2\nu})$. Hence one has from (5.9) that $w^*=O(\varepsilon^{\nu-1})$.

Thus by choice of $\nu=\max(N+2,4)$, the function w in T_ε and therefore Z is $O(\varepsilon^{N+1})$ uniformly on S_0. From this the assertion (5.1) follows.

Appendix. The fixed point of A_ε. In this appendix we sketch briefly a proof, for small ε, of the existence and uniqueness of a fixed point in $L_2(Q)$ of the nonlinear operator A_ε introduced in (5.8). The fixed point is shown also to be a classical solution of (5.5)–(5.7).

For several technical reasons involving the nonlinearity F, we were led to examine A_ε as a contractive map on the following closed subset $\mathfrak{S}$ of $L_2(Q)$ (as above $\|\cdot\|_p$ denotes the $L_p(Q)$ norm):

$$\mathfrak{S}:=\{v\in L_2(Q)\cap L_\infty(Q):\|v\|_2\le\rho,\ \|v\|_\infty\le\sigma\},$$

where ρ,σ (to be determined) are positive numbers in $(0,1)$ and may depend upon ε. Note that $\mathfrak{S}$ is not empty, since if

$$\|v\|_\infty\le[m(Q)]^{-1/2}\rho,$$

then $\|v\|_2\le\rho$, where $m(Q)$ denotes the area of Q and depends on ε.

Proceeding to the needed estimates, we see from the definition of G in (5.4) and Taylor's theorem,

$$|\bar{G}(v)(x,t)|\le|\zeta(x)||v(x,t)|^2\int_0^1\|F''\|_*\eta\,d\eta\le\frac{\|F''\|_*}{2}|v(x,t)|^2, \tag{A.1}$$

where

$$\|F^{(j)}\|_*=\sup_\lambda|F^{(j)}(\lambda)|$$

for $|\lambda| \leq (|U(x,t;\varepsilon)+\varepsilon V(x,t;\varepsilon)+X(x,t;\varepsilon)|+1)$, for (x,t) in S_0 and ε in $(0,\varepsilon_0]$. Thus $\|F^{(j)}\|_*$ depends on ε_0 determined in §3 but not on ε in the interval $(0,\varepsilon_0]$. The fact that $\|v\|_\infty \leq 1$ ensures that the argument of F'' lies on a bounded interval. If F'' was uniformly bounded on $\mathbb{R}$, there would be no need here to consider L_∞ as well as L_2.

Hence from (A.1)

$$\|\bar{G}(v)\|_\infty \leq \frac{\|F''\|_*}{2}\|v\|_\infty^2,$$

and similarly,

$$\|\bar{G}(v)\|_2 \leq \frac{\|F''\|_*}{2}\sigma\|v\|_2. \tag{A.2}$$

In a similar way for v_i in $\mathcal{S}$, $i=1,2$,

$$\|\bar{G}(v_1)-\bar{G}(v_2)\|_2 \leq \mathcal{C}(F)\big(\|v_1\|_\infty+\|v_2\|_\infty\big)\|v_1-v_2\|_2, \tag{A.3}$$

where $\mathcal{C}$ depends only on $\|F''\|_*$, $\|F'''\|_*$, and therefore on ε_0 but not ε.

The operator Λ_ε $(=\tilde{L}_\varepsilon^{-1})$ of §4 maps from $L_2(Q)$ into $\tilde{H}^1(Q)$, and $H^1(Q)$ is compactly imbedded into $L_2(Q)$ (see [13, p. 293]). In the present case the relevant estimate is

$$\|v\|_2 \leq [m(Q)]^{1/2}\|v\|_{H^1(Q)}$$

for any v in $\tilde{H}^1(Q)$. Therefore using the estimate (3.3) and (A.2)

$$\|A_\varepsilon(v)\|_2 \leq [m(Q)]^{1/2}[\lambda(\varepsilon)]^{-1/2}\sqrt{t_0}\cdot\left\{\frac{\|F''\|_*}{2}\sigma\|v\|_2+\|\bar{h}\|_2\right\}.$$

Similarly

$$\|A_\varepsilon(v)\|_\infty \leq [\lambda(\varepsilon)]^{-1/2}[m(Q)]^{1/2}\left\{\frac{\|F''\|_*}{2}\|v\|_\infty^2+\|\bar{h}\|_\infty\right\}.$$

From these estimates it follows that for small ε the operator A_ε maps $\mathcal{S}$ into itself, provided that $\sigma=\sigma(\varepsilon)$ and $\rho=\rho(\varepsilon)$ are chosen appropriately. Here we have used the fact that $\|\bar{h}\|_2$ and $\|\bar{h}\|_\infty$ are $O(\varepsilon^m)$ for as large an m as we wish by choice of ν.

Moreover by use of (A.3)

$$\|A_\varepsilon(v_1)-A_\varepsilon(v_2)\|_2 \leq \sqrt{t_0}\,\mathcal{C}(F)[m(Q)]^{1/2}[\lambda(\varepsilon)]^{-1/2}\cdot\big(\|v_1\|_\infty+\|v_2\|_\infty\big)\|v_1-v_2\|_2,$$

allowing us to choose ρ to meet the previous requirements and also to make A_ε a contraction on balls of radius ρ about the origin in $L_2(Q)$.

Thus A_ε has the desired fixed point w^* in $L_2(Q)$. But since

$$w^* = \Lambda_\varepsilon\bar{G}(w^*)+\Lambda_\varepsilon\bar{h},$$

it follows from the properties of Λ_ε of §4 that w^* is in $\tilde{H}^1(Q)$ and so by boot-strapping turns out to be a classical solution of (5.5)–(5.7), provided that F,f and g satisfy the restrictions of Theorem 1 of §5.

REFERENCES

[1] A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU, *Perturbations et "Augmentation" des conditions initiales*, Lecture Notes in Mathematics, 594, Springer-Verlag, Berlin, 1977, pp. 10–29.

[2] L. BOBISUD, *Degeneration of the solutions of certain well-posed systems of partial differential equations depending on a small parameter*, J. Math. Anal. Appl., 16 (1966), pp. 419–454.

[3] H. BRÉZIS AND L. NIRENBERG, *Characterization of the ranges of some nonlinear operators and applications to boundary value problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (IV), 5 (1978), pp. 225–326.

[4] L. CESARI AND R. KANNAN, *Existence of solutions of nonlinear hyperbolic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (IV), 6 (1979), pp. 573–592.

[5] B. F. ESHAM, *Singular perturbation problem for nonlinear evolution equations in Hilbert space*, to appear.

[6] R. GEEL, *Linear initial value problems with a singular perturbation of hyperbolic type*, Proc. Roy. Soc. Edinburgh A, 87 (1981), pp. 167–187.

[7] R. GEEL AND E. M. DE JAGER, *Hyperbolic singular perturbations of nonlinear first order differential equations*, Differential Equations and Applications, North-Holland, Amsterdam, 1978.

[8] J. GENET AND M. MADAUNE, *Singular perturbations for a class of nonlinear hyperbolic-hyperbolic problems*, J. Math. Anal. Appl., 64 (1978), pp. 1–24.

[9] G. C. HSIAO AND R. J. WEINACHT, *A singularly perturbed Cauchy problem*, J. Math. Anal. Appl., 71 (1979), pp. 242–250.

[10] G. C. HSIAO AND R. J. WEINACHT, *Singular perturbations for a weakly nonlinear hyperbolic equation*, Appl. Anal. 10 (1980), pp. 221–229.

[11] V. A. IL'IN, *On solvability of mixed problems for hyperbolic and parabolic equations*, Uspekhi Mat. Nauk, 15, 2(92), (1960), pp. 97–154; Russian Math. Surveys, 15, 2 (1960), pp. 85–142.

[12] E. M. DE JAGER, *Singular perturbations of hyperbolic type*, Nieuw Archief voor Wiskunde (3), Vol. XXIII (1975), pp. 145–171.

[13] A. KUFNER, O. JOHN AND S. FUCIK, *Function Spaces*, Noordhoff, Leyden, 1977.

[14] V. P. MIKHAILOV, *Partial Differential Equations*, MIR, Moscow, 1978. (In Russian.)

[15] C. S. MORAWETZ, *A uniqueness theorem in Frankl's problem*, Comm. Pure Appl. Math., 7 (1954), pp. 697–703.

[16] J. L. NOWINSKI, *Theory of Thermoelasticity with Applications*, Sijthoff and Noordhoff, Alphen Aan Den Rijn, the Netherlands, 1978.

[17] M. H. PROTTER, *Uniqueness theorems for the Tricomi problem*, J. Rat. Mech. Anal., 2 (1953), pp. 107–114, 4 (1955), pp. 721–732.

[18] D. R. SMITH, *The multivariable method in singular perturbation analysis*, SIAM Rev., 17 (1975), pp. 221–273.

[19] M. I. VISIK AND L. A. LYUSTERNIK, *Regular degeneration and boundary layer for linear differential equations with small parameter*, Uspehi Mat. Nauk, 12 (1957), pp. 3–122; Amer. Math. Soc. Transl. Ser. 2, 20 (1960), pp. 239–364.

[20] M. ZLAMAL, *On the mixed boundary value problem for a hyperbolic equation with a small parameter*, Czechoslovak Math. J., 9(84) (1959), pp. 218–242. (In Russian.)

[21] M. ZLAMAL, *On a singular perturbation problem concerning hyperbolic equations*, Lecture series 45, The Institute for Fluid Dynamics and Applied Mathematics, Univ. Maryland, College Park, MD, November, 1964.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0010 $01.25/0

A CHARACTERIZATION OF THE SPACES $\mathcal{S}_{1/k+1}^{k/k+1}$ BY MEANS OF HOLOMORPHIC SEMIGROUPS*

S. J. L. VAN EIJNDHOVEN,† J. DE GRAAF† AND R. S. PATHAK†

Abstract. The Gel'fand–Shilov spaces $\mathcal{S}_\alpha^\beta$, $\alpha=1/(k+1)$, $\beta=k/(k+1)$, are special cases of a general type of test function spaces introduced by de Graaf. We give a self-adjoint operator so that the test functions in those $\mathcal{S}_\alpha^\beta$ spaces can be expanded in terms of the eigenfunctions of that self-adjoint operator.

AMS-MOS subject classification (1980). Primary 46F05, 35K15

1. Introduction. De Bruijn's theory of generalized functions based on a specific one-parameter semigroup of smoothing operators [1] was generalized considerably by de Graaf [4]. In brief this extended theory can be described as follows: In a Hilbert space $\mathcal{X}$ consider the evolution equation

$$\frac{du}{dt}=-\mathfrak{A}u \tag{1.1}$$

where $\mathfrak{A}$ is a positive, self-adjoint operator, which is unbounded in order that the semigroup $(e^{-t\mathfrak{A}})_{t\geq 0}$ is smoothing. A solution u of (1.1) is called a trajectory if u satisfies

$$\forall t>0\ \forall \tau>0:\quad e^{-\tau\mathfrak{A}}u(t)=u(t+\tau), \tag{1.2i}$$

$$\forall t>0:\quad u(t)\in\mathcal{X}. \tag{1.2ii}$$

The limit $\lim_{t\downarrow 0}u(t)$ does not necessarily exist in $\mathcal{X}$!

The complex vector space of all trajectories is denoted by $\mathcal{T}_{\mathcal{X},\mathfrak{A}}$. The elements of $\mathcal{T}_{\mathcal{X},\mathfrak{A}}$ are called generalized functions.

The test function space $\mathcal{S}_{\mathcal{X},\mathfrak{A}}$ is the dense linear subspace of $\mathcal{X}$ consisting of smooth elements of the form $e^{-t\mathfrak{A}}h$, where $h\in\mathcal{X}$ and $t>0$; we have $\mathcal{S}_{\mathcal{X},\mathfrak{A}}=\cup_{t>0}e^{-t\mathfrak{A}}(\mathcal{X})$. The densely defined inverse of $e^{-t\mathfrak{A}}$ is denoted by $e^{t\mathfrak{A}}$. For each $\varphi\in\mathcal{S}_{\mathcal{X},\mathfrak{A}}$ there exists $\tau>0$ such that $e^{\tau\mathfrak{A}}\varphi$ makes sense. The pairing between $\mathcal{S}_{\mathcal{X},\mathfrak{A}}$ and $\mathcal{T}_{\mathcal{X},\mathfrak{A}}$ is defined by

$$\langle\varphi,F\rangle=:(e^{\tau\mathfrak{A}}\varphi,F(\tau)),\quad \varphi\in\mathcal{S}_{\mathcal{X},\mathfrak{A}},\quad F\in\mathcal{T}_{\mathcal{X},\mathfrak{A}}. \tag{1.3}$$

Here $(\cdot,\cdot)$ denotes the inner product in $\mathcal{X}$. Definition (1.3) makes sense for $\tau>0$ sufficiently small, and due to the trajectory property (1.2i) it does not depend on the specific choice of τ. For further results concerning this theory we refer to [4].

The aim of the present paper is to show that for certain Gel'fand–Shilov spaces $\mathcal{S}_\alpha^\beta$ [2] there exists an operator $\mathfrak{A}$ such that $\mathcal{S}_\alpha^\beta=\mathcal{S}_{\mathcal{X},\mathfrak{A}}$. This leads to the result that the elements of the dual of $\mathcal{S}_\alpha^\beta$ can be interpreted as trajectories. Furthermore, we find that a function in the studied $\mathcal{S}_\alpha^\beta$-spaces can be developed in a series of certain orthonormal functions.

*Received by the editors February 19, 1982.

†Department of Mathematics, Technological University, Eindhoven, the Netherlands. One of the authors (SJLVE) was supported by a grant from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

2. Eigenfunction expansions of test functions in $\mathcal{S}_\alpha^\beta$. Let us consider the following eigenvalue problem in $\mathcal{L}_2(\mathbb{R})$:

$$\frac{d^2}{dx^2}y+(\lambda-x^{2k})y=0, \tag{2.1}$$

where λ is a real number and k a positive integer. It is well-known that the operator $-d^2/dx^2+x^{2k}$ has a point spectrum and the set of eigenvalues (λ_n) is real, positive and unbounded. In the sequel we shall regard it as ordered with $\lambda_{n+1}\geq\lambda_n$, $n=0,1,,\cdots$. The corresponding normalized eigenfunctions $\{\psi_n\}$ form a complete orthonormal basis in $\mathcal{L}_2(\mathbb{R})$. So by the Riesz–Fischer theorem every $f\in\mathcal{L}_2(\mathbb{R})$ can be represented by

$$f=\sum_{n=0}^{\infty}a_n\psi_n, \tag{2.2}$$

where $a_n=(f,\psi_n)$ is an $\mathcal{L}_2$-sequence.

First of all we gather some of the estimates for the eigenvalues λ_n and the eigenfunctions ψ_n of the problem (2.1), and then characterize $\{\psi_n\}$ as elements of certain $\mathcal{S}_\alpha^\beta$-spaces. We take $\psi_n(x)>0$ for large positive values of x, cf. Titchmarsh [5, Chap. VIII].

From Titchmarsh [5, p. 144] we have

$$\lambda_n=O(n^{2k/(k+1)}),\qquad n\to\infty. \tag{2.3}$$

According to Titchmarsh we have the following estimates for the normalized eigenfunctions

$$|\psi_n(x)|\leq\frac{2}{3}\lambda_n^{1+3/4k}\quad\text{for all } x\in\mathbb{R},\ n\in\mathbb{N}\quad[5,\text{p. }168], \tag{2.4}$$

$$|\psi_n(x)|\leq\psi_n(x_0)\exp\left\{-\int_{x_0}^{x}(u^{2k}-\lambda_n)^{1/2}du\right\}\quad\text{for } x\geq x_0\geq\lambda_n^{k/2}\quad[5,\text{p. }165]. \tag{2.5}$$

We take $x_0=(\frac{4}{3}\lambda_n)^{1/2k}$. From a straightforward calculation it follows that

$$|\psi_n(x)|\leq\frac{2}{3}\lambda_n^{1+3/4k}\exp\left\{-\frac{1}{4}\frac{1}{k+1}|x|^{k+1}\right\}$$

for $|x|\geq2\lambda_n^{1/2k}$. For any number a, $0<a<1/4(k+1)$, we have

$$|\psi_n(x)|\leq K_n\exp\left(-a|x|^{k+1}\right),\qquad x\in\mathbb{R}, \tag{2.6}$$

where

$$K_n=\frac{2}{3}\lambda_n^{1+3/4k}\exp\left(2^{k+1}a\lambda_n^{(k+1)/2k}\right).$$

The eigenfunction $\psi_n(x)$ can be extended to an entire function $\psi_n(z)$. We want to estimate $\psi_n(z)$ in the complex plane. First we produce an estimate for $|\psi_n'(0)|$. Let $\xi>0$ denote a point at which ψ_n^2 reaches its absolute maximum. We have $0\leq\xi\leq\lambda_n^{1/2k}$. Integrate the equality

$$-\frac{d}{dx}(\psi_n')^2=(\lambda_n-x^{2k})\frac{d}{dx}(\psi_n^2)$$

from 0 to ξ. A crude estimate yields

$$|\psi_n'(0)| \leq \frac{2}{3}\sqrt{1+2k}\,\lambda_n^{1+3/2+3/4k}.$$

Next, following the technique of Titchmarsh [5, p. 172] it can be shown that

$$\psi_n(z) = y^{(0)}(z) + \sum_{m=1}^{\infty} \{y^{(m)}(z) - y^{(m-1)}(z)\}, \qquad z \in \mathbb{C}.$$

Here $y^{(0)}(z) = \psi_n(0) + z\psi_n'(0)$ and $y^{(m)}(z)$, $m \geq 1$, can be obtained from

$$y^{(m)}(z) = y^{(0)}(z) + \int_0^z (s^{2k} - \lambda_n) y^{(m-1)}(s)(w-s)\,ds.$$

With

$$|y^{(m)}(z) - y^{(m-1)}(z)| \leq |y^{(0)}(z)| \{|z|^{2k} + \lambda_n\}^m \frac{|z|^{2m}}{(2m)!}$$

we get the estimate

$$|\psi_n(z)| \leq K_n(|z|)\exp\left(|z|^{k+1} + \lambda_n^{1/2}|z|\right).$$

Here

$$K_n(|z|) = \frac{4}{3}\lambda_n^{1+3/4k}\left(1 + (1+2k)^{1/2}\lambda_n^{1/2}|z|\right) \geq |y^{(0)}(z)|.$$

Now let $d > 0$. Then

$$\exp\left(\lambda_n^{1/2}|z|\right) \leq \exp\left(d^{-k}|z|^{k+1}\right)$$

whenever $|z| \geq d\lambda_n^{1/2k}$ and

$$\exp\left(\lambda_n^{1/2}|z|\right) \leq \exp\left(d\lambda_n^{(k+1)/2k}\right)$$

whenever $|z| \leq d\lambda_n^{k/2}$. Thus we have

$$|\psi_n(z)| \leq K_n(|z|)\exp\left(d\lambda_n^{(k+1)/2k}\right)\exp(1 + d^{-k})|z|^{k+1}. \tag{2.7}$$

THEOREM 1. *The eigenfunctions ψ_n of the eigenvalue problem* (2.1) *are elements of the space $\mathcal{S}_\alpha^\beta$, where $\alpha = 1/(k+1)$ and $\beta = k/(k+1)$.*

Proof. Since ψ_n is an entire function and since it satisfies (2.6) and (2.7), in view of the criterion of Gel'fand and Shilov [2, p. 220], the result follows. □

THEOREM 2. *Let $f \in \mathcal{L}_2(\mathbb{R})$,*

$$f = \sum_{n=0}^{\infty} a_n \psi_n,$$

and suppose there is $\tau > 0$ such that

$$a_n = O\left(\exp\left(-\tau\lambda_n^{(k+1)/2k}\right)\right).$$

Then $f \in \mathcal{S}_{1/k+1}^{k/k+1}$.

Proof. In (2.6) we can take $a>0$ so small that $\tau>a2^{k+1}$. Then for some $C>0$ and all $x\in\mathbb{R}$

$$|f(x)|\leq \sum_{n=0}^{\infty} |a_n||\psi_n(x)|$$

$$\leq C\sum_{n=0}^{\infty} K_n\exp\{-(\tau-a2^{k+1})\lambda_n^{(k+1)/2k}\}\exp\bigl(-a|x|^{k+1}\bigr).$$

So $|f(x)|\leq C'\exp(-a|x|^{k+1})$ for some $C'>0$. Further we can take $d>0$ and $d<\tau$, so that with the aid of (2.7)

$$|f(z)|\leq \sum_{n=0}^{\infty} |a_n||\psi_n(z)|$$

$$\leq \exp\bigl((1+d^{-k})|z|^{k+1}\bigr)\sum_{n=0}^{\infty} K_n(|z|)\exp\bigl(-(\tau-d)\lambda_n^{(k+1)/2k}\bigr)$$

$$\leq C''\exp\bigl((1+d^{-k})|z|^{k+1}\bigr)$$

for some $C''>0$. By the criterion of Gel'fand and Shilov as used in the proof of Theorem 1, $f\in\mathcal{S}_{1/k+1}^{k/k+1}$. □

Let $\mathfrak{A}_k$ be the self-adjoint operator in $\mathcal{L}_2(\mathbb{R})$ defined by

$$\mathfrak{A}_k=-\frac{d^2}{dx^2}+x^{2k}. \tag{2.8}$$

Then as a corollary of Theorem 2 we have

COROLLARY 1. *The test function space* $\mathcal{S}_{\mathcal{L}_2(\mathbb{R}),\mathfrak{B}_k}$ *is included in* $\mathcal{S}_{1/k+1}^{k/k+1}$. *Here* $\mathfrak{B}_k=(\mathfrak{A}_k)^{(k+1)/2k}$.

Proof. The functions ψ_n are the eigenfunctions of the positive self-adjoint operator $\mathfrak{B}_k$ with eigenvalues $\lambda_n^{(k+1)/2k}$. Let $f\in\mathcal{S}_{\mathcal{L}_2(\mathbb{R}),\mathfrak{B}_k}$. Then there exists $h\in\mathcal{L}_2(\mathbb{R})$ and $\tau>0$ such that

$$f=e^{-\tau\mathfrak{B}_k}h.$$

This provides $(f,\psi_n)=\exp(-\tau\lambda_n^{(k+1)/2k})(h,\psi_n)$. So the coefficients (f,ψ_n) are of the order $\exp(-\tau\lambda_n^{(k+1)/2k})$. By Theorem 2 we have $f\in\mathcal{S}_{1/k+1}^{k/k+1}$. □

We want to prove the converse of Corollary 1:

THEOREM 3.

$$\mathcal{S}_{\mathcal{L}_2(\mathbb{R}),\mathfrak{B}_k}=\mathcal{S}_{1/k+1}^{k/k+1}.$$

In the proof of this theorem we need some lemmas.

LEMMA 1. *Let* i_r, j_r *be nonnegative integers for* $r=1,2,\cdots,n$. *Then*

$$\mathbf{D}^{i_1}x^{j_1}\mathbf{D}^{i_2}\cdots\mathbf{D}^{i_n}x^{j_n}=\sum_{l\in\mathbb{N}^n} c_{ij}(l)x^{|j-l|}\mathbf{D}^{|i-l|},$$

where $\mathbf{D}$ *is the differential operator* d/dx *and where the coefficients* $c_{ij}(l)$ *satisfy*

$$|c_{ij}(l)|\leq\frac{1}{l!}\,\frac{j!}{(j-l)!}\,\frac{|i|!}{|i-l|!}$$

($c_{ij}(l)=0$ *if* $l>\min(i,j)$).

We use multi-indices, and $|i|=i_1+i_2+\cdots+i_n$, $i!=i_1!i_2!\cdots i_n!$, etc.

Proof. See Goodman [3, p. 67]. □

LEMMA 2. *Let f be an infinitely differentiable function which satisfies the following inequalities for fixed $A,B,C>0$ and $\alpha,\beta>0$, $\alpha+\beta\geq 1$:*

$$|(x^k\mathbf{D}^l f)(x)|\leq CA^kB^l k^{\alpha k}l^{\beta l}, \qquad k,l=0,1,2,\cdots. \tag{2.9}$$

Then for each $n\in\mathbb{N}$ and $i,j\in\mathbb{N}^n$

$$|(\mathbf{D}^{i_1}x^{j_1}\cdots\mathbf{D}^{i_n}x^{j_n}f)(x)|\leq C_1A_1^{|j|}B_1^{|i|}|j|^{\alpha|j|}|i|^{\beta|i|}$$

where $C_1=C$, $A_1=2^{\sigma\beta+1}e^{\sigma\alpha}A$, $B_1=2^{\sigma\alpha}e^{\sigma\beta}B$ and $\sigma=(\alpha+\beta)^{-1}$.

Proof. Let $n\in\mathbb{N}$ and $i,j\in\mathbb{N}^n$. Then by Lemma 1

$$|(\mathbf{D}^{i_1}x^{j_1}\cdots\mathbf{D}^{i_n}x^{j_n})f(x)|\leq\sum_l|c_{ij}(l)||(x^{|j-l|}\mathbf{D}^{|i-l|}f(x))|.$$

With the assumption (2.9) we estimate this series as follows:

$$\begin{aligned}&\sum_{l\leq\min(i,j)}|c_{ij}(l)||(x^{|j-l|}\mathbf{D}^{|i-l|}f)(x)|\\&\quad\leq C\sum_l\frac{1}{l!}\frac{j!}{(j-l)!}\frac{|i|!}{|i-l|!}A^{|j-l|}B^{|i-l|}|j-l|^{\alpha|j-l|}|i-l|^{\beta|i-l|}\\&\quad\leq CA^{|j|}B^{|i|}\sum_l\frac{1}{l!}\frac{j!}{(j-l)!}\frac{|i|!}{|i-l|!}|j-l|^{\alpha|j-l|}|i-l|^{\beta|i-l|}.\end{aligned}$$

The latter series can be treated as follows

$$\begin{aligned}&\sum_{l\leq\min(i,j)}\frac{1}{l!}\frac{j!}{(j-l)!}\frac{|i|!}{|i-l|!}|j-l|^{\alpha|j-l|}|i-l|^{\beta|i-l|}\\&\quad\leq\sup_{|l|\leq|i|}\frac{|i|!|i-l|^{\beta|i-l|}}{|i-l|!(|l|!)^{\sigma\alpha}}\sup_{|l|\leq|j|}\left(\frac{|j|!|j-l|^{\alpha|j-l|}}{|j-l|!(|l|!)^{\sigma\beta}}\right)\sum_{l\leq j}\binom{j}{l}\binom{|j|}{|l|}^{-1}.\end{aligned}$$

We have

$$\sum_{l\leq j}\binom{j}{l}\binom{|j|}{|l|}^{-1}\leq\sum_{l\leq j}\binom{j}{l}=2^{|j|}.$$

With the aid of the inequality $n!<n^n<n!e^n$:

$$\begin{aligned}\left(\frac{|i|!|i-l|^{\beta|i-l|}}{|i-l|!(|l|!)^{\sigma\alpha}}\right)&\leq\binom{|i|}{|l|}^{\sigma\alpha}\frac{(|i|!)^{\sigma\beta}}{(|i-l|!)^{\sigma\beta}}|i-l|^{\beta|i-l|}\\&\leq 2^{\sigma\alpha|i|}e^{\sigma\beta|i|}(|i|!)^{\sigma\beta}(|i-l|)^{(1-\sigma)\beta|i-l|}\leq 2^{\sigma\alpha|i|}e^{\sigma\beta|i|}|i|^{\beta|i|}\end{aligned}$$

and similarly

$$\left(\frac{|j|!|j-l|^{\alpha|j-l|}}{|j-l|!(|l|!)^{\sigma\beta}}\right)\leq 2^{\sigma\beta|j|}e^{\sigma\alpha|j|}|j|^{\alpha|j|}.$$

Combining these results, we derive

$$\left|(\mathbf{D}^{i_1}x^{j_1}\cdots\mathbf{D}^{i_n}x^{j_n}f)(x)\right|\leq CA_1^{|j|}B_1^{|i|}|j|^{\alpha|j|}|i|^{\beta|i|},$$

where $A_1=2^{\sigma\beta+1}e^{\sigma\alpha}A$, $B_1=2^{\sigma\alpha}e^{\sigma\beta}B$. □

LEMMA 3. *For $f\in\mathcal{S}_{1/k+1}^{k/k+1}$ we have*

$$\left|(\mathbf{D}^2-x^{2k})^p f(x)\right|\leq KN^p p^{2pk/(k+1)},\qquad p=0,1,2,\cdots,$$

where K and N are fixed positive constants depending on f.

Proof. Let $\alpha=1/(k+1)$, $\beta=k/(k+1)$. Let $f\in\mathcal{S}_\alpha^\beta$. Then there are positive constants A,B,C such that for all $x\in\mathbb{R}$

$$\left|(x^l\mathbf{D}^q f)(x)\right|\leq CA^lB^q l^{\alpha l}q^{\beta q},$$

with $l,q=0,1,2,\cdots$.

Now let $p\in\mathbb{N}$. Then

$$(\mathbf{D}^2-x^{2k})^p=\sum_{s=0}^{p}V_s(\mathbf{D}^2,x^{2k}),$$

where $V_s(\mathbf{D}^2,x^{2k})$ consists of a sum of $\binom{p}{s}$ combinations of the form

$$(\mathbf{D}^2)^{i_1}(x^{2k})^{j_1}\cdots(\mathbf{D}^2)^{i_n}(x^{2k})^{j_n}$$

where $i_1+\cdots+i_n=s$ and $j_1+\cdots+j_n=p-s$. With the aid of Lemma 2 we have

$$\left|V_s(\mathbf{D}^2,x^{2k})f(x)\right|\leq\binom{p}{s}CA_1^{2k(p-s)}B_1^{2s}(2k(p-s))^{2\alpha k(p-s)}(2s)^{2\beta s},$$

with $A_1=2^{\beta+1}e^{\alpha}A$ and $B_1=2^{\alpha}e^{\beta}B$. So

$$\begin{aligned}\left|(\mathbf{D}^2-x^{2k})^p f(x)\right|&\leq C\sum_{s=0}^{p}\binom{p}{s}A_2^{p-s}B_2^s(p-s)^{2\alpha k(p-s)}s^{2\beta s}\\&\leq C\sum_{s=0}^{p}\binom{p}{s}A_2^{p-s}B_2^s(p^{2\alpha k})^{p-s}(p^{2\beta})^s\\&=C(A_2p^{2\alpha k}+B_2p^{2\beta})^p.\end{aligned}$$

Substituting the values of α and β it follows that

$$\left|(\mathbf{D}^2-x^{2k})^p f(x)\right|\leq C(A_2+B_2)^p p^{2pk/(k+1)}$$

where $A_2=((2k)^\alpha A_1)^{2k}$ and $B_2=2^\beta B_1$. □

Proof of Theorem 3. Because of Corollary 1 we only have to prove the inclusion

$$\mathcal{S}_{1/k+1}^{k/k+1}\subset\mathcal{S}_{\mathcal{L}_2(\mathbb{R}),\mathfrak{B}_k}.$$

So let $f\in\mathcal{S}_{1/k+1}^{k/k+1}$. Put $a_n=(f,\psi_n)$, $n\in\mathbb{N}$. Then for each $p\in\mathbb{N}$ fixed

$$a_n=(f,\psi_n)=\lambda_n^{-p}\left((-\mathbf{D}^2+x^{2k})^p f,\psi_n\right).$$

With the aid of Lemma 3 we get positive constants K_f and N_f such that

$$\left\|(-\mathbf{D}^2+x^{2k})^p f\right\|_\infty\leq K_f N_f^p p^{2pk/(k+1)}.$$

And

$$|a_n| \leq \lambda_n^{-p} \left\| (-\mathbf{D}^2 + x^{2k})^p f \right\|_\infty \|\psi_n\|_1, \qquad n = 0, 1, 2, \cdots.$$

By (2.4) and (2.5)

$$\|\psi_n\|_1 = \int_{-\infty}^{\infty} |\psi_n(x)|\, dx = \left(\int_{|x| \leq 2\lambda_n^{k/2}} + \int_{|x| > 2\lambda_n^{k/2}} \right) |\psi_n(x)|\, dx$$

$$\leq \frac{8}{3} \lambda_n^{1+5/4k} + c_k \lambda_n^{1+3/4k}$$

where c_k only depends on k. Therefore

$$|a_n| \leq c_k' \lambda_n^{1+5/4k} \lambda_n^{-p} K_f N_f^p p^{2pk/(k+1)}.$$

Finally taking the infinum of the right-hand side with respect to p we arrive at

$$|a_n| \leq c_k' K_f \lambda_n^{1+5/4k} \exp - \left\{ 2\beta e^{-1} N_f^{-1/2\beta} \right\} \lambda_n^{1/2\beta}$$

with $\beta = k/(k+1)$. From this the assertion follows. □

By taking Fourier transforms in Theorem 3 we derive easily

THEOREM 4.

$$\mathcal{S}_{k/k+1}^{1/k+1} = \mathcal{S}_{\mathcal{L}_2(\mathbb{R}), \tilde{\mathfrak{B}}_k},$$

where $\tilde{\mathfrak{B}}_k = ((-d^2/dx^2)^k + x^2)^{(k+1)/2k}$.

REFERENCES

[1] N. G. DE BRUIJN, *A theory of generalized functions with applications to Wigner distributions and Weyl correspondence*, Nieuw Archief voor Wiskunde, 3, XX1 (1973), pp. 205–280.

[2] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions*, Vol. 2, Academic Press, New York, 1968.

[3] R. GOODMAN, *Analytic and entire vectors for representations of Lie groups*, Trans. Amer. Math. Soc., 143 (1969), pp. 55–76.

[4] J. DE GRAAF, *A theory of generalized functions based on holomorphic semigroups*, T.H.-Report 79-WSK-02, Eindhoven University of Technology, Eindhoven, the Netherlands, 1979.

[5] E. C. TITCHMARSH, *Eigenfunction Expansions Associated with Second Order Differential Equations, Part* I, Oxford Univ. Press, Oxford, 1962.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406-0011 $01.25/0

ON SOME INTEGRAL EQUATIONS WITH LOCALLY FINITE MEASURES AND L^∞-PERTURBATIONS*

STIG-OLOF LONDEN†

Abstract. Let $g\in C(R)$, $f\in L^\infty_{\text{loc}}(R^+)$ and let μ be a real locally finite positive definite Borel measure on R^+. We investigate a relation between the solution of the nonlinear scalar Volterra equation

$$x'(t)+\int_{[0,t]} g(x(t-s))\,d\mu(s)=f(t), \qquad t\in R^+, \quad x(0)=x_0,$$

and the solution of the linear equation with the same data

$$z'(t)+\int_{[0,t]} z(t-s)\,d\mu(s)=f(t), \qquad t\in R^+, \quad z(0)=x_0.$$

This relation, when combined with results (established in this paper) on the set of bounded solutions of certain limit equations

$$y(t)+\int_{R^+} g(y(t-s))a(s)\,ds=0, \qquad t\in R,$$

allows us to obtain new asymptotic results for $x(t)$ in the case when both μ and f are large in a precise sense.

1. Introduction. In this paper we analyze a certain connection between the asymptotic behavior of the solutions of the nonlinear scalar Volterra equation

$$x'(t)+\int_{[0,t]} g(x(t-s))\,d\mu(s)=f(t), \qquad t\in R^+=[0,\infty), \qquad x(0)=x_0, \tag{1.1}$$

and the corresponding behavior of the solution of the linear equation with the same data

$$z'(t)+\int_{[0,t]} z(t-s)\,d\mu(s)=f(t), \qquad t\in R^+, \quad z(0)=x_0. \tag{1.2}$$

As a consequence of this connection we obtain some new asymptotic results on (1.1) in the case when both μ and f are large.

In the equations above, g, μ, f, x_0 are given real-valued, while x, z stand for the solutions. These solutions are always assumed to exist for $t\in R^+$, to be locally bounded, and to satisfy the corresponding equations almost everywhere on R^+. Throughout the article the following basic hypotheses on g, μ, f will be made:

(1.3)
(i) $g\in C(R)$,
(ii) μ is a real, locally finite, positive definite measure on R^+,
(iii) $f\in L^1_{\text{loc}}(R^+)$.

Define $Q(\varphi,\mu,T)$ for $\varphi\in L^2_{\text{loc}}(R^+)$, $T>0$, by

(1.4)
$$Q(\varphi,\mu,T)=\int_0^T \varphi(t)(\varphi*\mu)(t)\,dt, \quad \text{where}$$
$$(\varphi*\mu)(t)\overset{\text{def}}{=}\int_{[0,t]} \varphi(t-s)\,d\mu(s),$$

*Received by the editors June 11, 1981, and in revised form May 1, 1982. This work was sponsored by the United States Army under contract DAAG29-80-C-0041. The work was done while the author was visiting the Department of Mathematics and the Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

†Institute of Mathematics, Helsinki University of Technology, Espoo 15, Finland.

and let $x \in L^\infty(R^+)$. Then, as is well known [9], [10], a large amount of information concerning the asymptotic behavior of $g(x(t))$ can be obtained provided one succeeds in establishing

$$\sup_{T>0} Q(g,\mu,T) < \infty, \tag{1.5}$$

where $g=g(t)=g(x(t))$. Note that if in addition to $x \in L^\infty(R^+)$ one takes f small, i.e., $f \in L^1(R^+)$, then (1.5) follows easily.

If in addition to (1.3iii) f merely satisfies

$$\lim_{t\to\infty} f(t)=0, \tag{1.6}$$

then the asymptotic analysis of $x(t)$ becomes significantly more difficult, as (1.5) is now out of reach. However, by taking μ small enough, in particular by assuming

$$\int_{R^+} t\, d|\mu|(t) < \infty \tag{1.7}$$

($|\mu|$ is the total variation measure of μ), and by working with the limit equation corresponding to (1.1)

$$y'(t)+\int_{R^+} g(y(t-s))\, d\mu(s)=0, \qquad t\in R, \tag{1.8}$$

one may even now obtain asymptotic results on bounded solutions of (1.1) [3], [12]. Observe furthermore that if in addition to (1.3i) $g(x)$ is taken locally Lipschitzian, then (1.7) may be weakened to μ finite [4].

The aim of the present work was originally to extend the results of [3], [4], [12] so as to apply to equations with μ only locally finite without excluding the possibility that f satisfies only (1.6). However, making use of a simple device, one may connect the asymptotics of (1.1) and (1.2) and thus reduce (under certain hypotheses) the asymptotic analysis of (1.1) to that of (1.2). The fact that (1.2) can be explicitly solved for z independently of the size of μ and f then allows us to realize our original goal. This approach has earlier been applied in [2, Thm. 3] to obtain a result on the integrated version of (1.1). Our Theorem 1 is essentially a restatement for (1.1) of this result.

Theorem 1 has the advantage of having a short and lucid proof. Also observe the important point that nothing but continuity is imposed on g. The assumptions of Theorem 1 do, however, include a moment condition, (1.11), on the second derivative of the differential resolvent of μ. Although this condition is satisfied (Lemma 1 below) for $d\mu(t)=a(t)\,dt$ with $a(t)$ nonintegrable but sufficiently monotone, it is still the case that verification of (1.11) in general is quite hard if μ is only locally finite. It should also be observed that Theorem 1 requires $\hat{\mu}(\omega)$ to be finite for $\omega \neq 0$, thus excluding cases like $d\mu=a(t)\,dt$ with $a(t)=t^{-1/2}\cos t$.

One is consequently motivated to try to remove (1.11), (1.12). This is done in a series of steps, Theorems 2–4. Theorems 2 and 3 constitute auxiliary results, but as they are of independent interest we prefer to state them separately. Observe that these statements concern equations with a finite measure α. Theorem 4 then corresponds to Theorem 1 but (1.11), (1.12) are now absent from the assumptions. Certain other conditions have instead been added, in particular on $g(x)$. These additional assumptions on g have the advantage of being easily checked, and they are not overly restrictive. The added assumption that $g(x)$ be locally Lipschitzian is basic to the approach we use. The remaining additional hypotheses on g, roughly speaking, result from the fact that in the first part of the proof of Theorem 2 we establish $g(y(t)) \in L^2(R)$

(for which some condition of type (1.24) is needed if $0 \in Z$), and not $[y(t)+g(y(t))] \in L^2(R)$ (which perhaps only requires that g satisfies some smoothness condition). Although the latter conclusion undoubtedly is the natural one (under the assumptions on α made in Theorem 2), we have not been able to establish it without any sign condition on g.

Our last result, Theorem 5, states a new boundedness result on (1.1). It displays a connection between the existence of bounded solutions of (1.1) and the total variation of solutions of (1.2).

THEOREM 1. *Let* (1.3) *hold and assume* $r \in \mathrm{LAC}(R^+)$ *satisfies*

$$r'(t)+(r*\mu)(t)=0 \quad a.e.\ on\ R^+, \quad r(0)=1, \tag{1.9}$$

$$r' \in (L^1 \cap \mathrm{NBV})(R^+). \tag{1.10}$$

Define ν *to be the measure corresponding to* $-r'$, *thus* $\nu([0,t])=-r'(t)$, $t \geq 0$, *and let*

$$\int_{R^+} t\, d|\nu|(t) < \infty. \tag{1.11}$$

Suppose

$$|\hat{\mu}(\omega)| < \infty, \qquad \omega \neq 0, \tag{1.12}$$

and let the set Z *defined by* $Z=\{\omega \mid \omega \neq 0,\ \mathrm{Re}\,\hat{\mu}(\omega)=0\}$ *be at most denumerable and such that*

$$\mathrm{Im}\,\hat{\mu}(\omega)=0, \qquad \omega \in Z. \tag{1.13}$$

Finally let x, z *satisfy respectively* (1.1) *and* (1.2) *and be such that*

$$x \in (\mathrm{LAC} \cap L^\infty)(R^+), \qquad z \in \mathrm{LAC}(R^+). \tag{1.14}$$

Then if

$$\lim_{t \to \infty} z(t)=z(\infty) \tag{1.15}$$

exists (*and is finite*) *one has*

$$\lim_{t \to \infty} [x(t+d)-x(t)]=0 \quad \forall d>0, \tag{1.16}$$

$$\lim_{t \to \infty} [r(\infty)x(t)+[1-r(\infty)]\, g(x(t))]=z(\infty). \tag{1.17}$$

If in addition $z' \in L^\infty(R^+)$, $\lim_{t \to \infty} \operatorname{ess\,sup}_{s \geq t} |z'(s)|=0$ *then* $\lim_{t \to \infty} \operatorname{ess\,sup}_{s \geq t} |x'(s)|=0$.

By $\hat{\mu}(\omega)$, $\omega \neq 0$, we mean $\lim_{s \to i\omega,\, \mathrm{Re}\, s>0} \tilde{\mu}(s)$ where $\tilde{\mu}(s)=\int_{R^+} e^{-st}\, d\mu(t)$. To see that this is well defined, note at first that as μ is a positive definite measure then μ is a tempered distribution [9, p. 229], and so the Laplace transform $\tilde{\mu}(s)$ exists for $\mathrm{Re}\, s>0$. Then observe that by (1.9)

$$-\int_{R^+} e^{-st} r'(t)\, dt=\tilde{\mu}(s)[s+\tilde{\mu}(s)]^{-1}, \qquad \mathrm{Re}\, s>0.$$

By (1.10) the left side is continuous for $\mathrm{Re}\, s \geq 0$. Hence $\lim_{s \to i\omega,\, \mathrm{Re}\, s>0} \tilde{\mu}(s)[s+\tilde{\mu}(s)]^{-1}$ exists for $\omega \in R$. One concludes that $\lim_{s \to i\omega,\, \mathrm{Re}\, s>0} \tilde{\mu}(s)$ exists, possibly infinite, for $\omega \neq 0$. The assumption (1.12) does, however, exclude this last possibility.

Concerning (1.10) note that this condition is (locally with respect to ω near $\omega=0$) weaker than the assumption $r \in L^1(R^+)$. This is seen as follows. The Fourier transforms

$\hat{r}'$, $\hat{\nu}$ (for $h\in L^1(R)$, let $\hat{h}\overset{\text{def}}{=}\int_R e^{-i\omega t}h(t)\,dt$) may be written respectively as

$$-\frac{1}{i\omega[\hat{\mu}]^{-1}+1}, \qquad \frac{i\omega}{i\omega[\hat{\mu}]^{-1}+1}, \qquad \omega\neq 0.$$

Thus (1.10) requires (locally) $i\omega[\hat{\mu}]^{-1}$ to behave as the transform of an $L^1(R)$-function whereas the assumption $r\in L^1(R^+)$ imposes (locally) the same behavior on $\hat{\mu}^{-1}$. The former is clearly a weaker condition near $\omega=0$. If for example $d\mu(t)=da(t)$ with $a(0)=0$, $a(t)=1$, $0<t\leq 1$; $a(t)=0$, $t>1$; then $\hat{\mu}(\omega)=1-\cos\omega+i\sin\omega$ and thus $\hat{\mu}$ does locally near $\omega=0$ conform to the requirements imposed by Theorem 1. Yet $\hat{\mu}(0)=0$.

In applications, one of course frequently has $r(\infty)=0$. In this case (1.17) reduces to $\lim_{t\to\infty} g(x(t))=z(\infty)$.

A class of only locally finite positive definite measures for which the corresponding differential resolvents do satisfy (1.10), (1.11) is given by

LEMMA 1. *Let $d\mu=a(t)\,dt$ where $a(t)$ is nonnegative, nonincreasing and convex on R^+ with $a\in L^1(0,1)$ and $s+\tilde{a}(s)\neq 0$, $\mathrm{Re}\,s\geq 0$. Then $r\in L^1(R^+)$ and* (1.10) *hold. If in addition $-a'(t)$ is convex then* (1.11) *is satisfied.*

The fact that $r\in L^1(R^+)$ under the assumptions of Lemma 1 is proved in [8]. The assertions (1.10), (1.11) follow by straightforward estimates making use of [8, Lemma 1], [1, Lemma 5.1]. See [5] for more details.

Our next result constitutes a first step towards eliminating (1.11), (1.12) from the hypotheses of Theorem 1. It gives conditions under which the set of bounded solutions of the equation

$$y(t)+\int_{R^+} g(y(t-s))\alpha([0,s])\,ds=0, \qquad t\in R, \tag{1.18}$$

contains only the trivial solution, in case α is a finite measure and $\alpha(R^+)=0$.

Define $a(t)=\alpha([0,t])$. Thus $\hat{\alpha}(\omega)=\int_{R^+} e^{-i\omega t}\,d\alpha(t)$, and $\hat{a}(\omega)=\int_{R^+} e^{-i\omega t}a(t)\,dt$.

THEOREM 2. *Let*

$$g(x) \text{ be real, locally Lipschitzian, } x\in R, \tag{1.19}$$

$$\alpha \text{ be a real, finite, positive definite Borel measure on } R^+, \tag{1.20}$$

$$a\in L^1(R^+). \tag{1.21}$$

Define Z by $Z=\{\omega|\mathrm{Re}\,\hat{\alpha}(\omega)=0\}$ and suppose that Z can be written as the union of three pairwise disjoint sets Z_1, Z_2, $\{0\}$, such that

$$\mathrm{Im}\,\hat{\alpha}(\omega)=0, \qquad \omega\in Z_1, \tag{1.22}$$

$$\mathrm{Im}\,\hat{\alpha}(\omega)\neq 0, \quad \omega\in Z_2, \quad \inf_{\omega\in Z_2\cup\{0\}} \mathrm{Re}\,\hat{a}(\omega)>0. \tag{1.23}$$

Finally assume that for some $K>0$

$$xg(x)\geq 0, \qquad |x|\leq K. \tag{1.24}$$

Define $Y_K=\{y|y\in \mathrm{LAC}(R), y \text{ satisfies } (1.18), \|y\|_{L^\infty(R)}\leq K\}$. Then

$$Y_k=\{0\}. \tag{1.25}$$

Observe that if y satisfies (1.18) then y also satisfies

$$y'(t)+\int_{R^+} g(y(t-s))\,d\alpha(s)=0 \quad \text{a.e. on } R. \tag{1.26}$$

Also note that (1.21) and the second part of (1.23) imply that $Z_2\cup\{0\}$ must be compact.

From Theorem 2 one immediately deduces the following result concerning the asymptotic behavior of solutions of

$$x(t)+\int_{[0,t]} g(x(t-s))\alpha([0,s])\,ds=F(t), \qquad t\in R^+. \tag{1.27}$$

THEOREM 3. *Let g, α be as in Theorem 2 and suppose F is such that*

$$F\in \mathrm{BC}(R^+), \qquad \lim_{t\to\infty} F(t)=0. \tag{1.28}$$

Let x be the solution of (1.27) *and assume*

$$\|x\|_{L^\infty(R^+)}\le K, \tag{1.29}$$

where K is as in (1.24). *Then*

$$\lim_{t\to\infty} x(t)=0, \qquad \lim_{t\to\infty} g(x(t))=0. \tag{1.30}$$

If in addition,

$$F\in \mathrm{LAC}(R^+), \qquad \lim_{t\to\infty} \operatorname{ess\,sup}_{s\ge t}|F'(s)|=0, \tag{1.31}$$

then

$$x\in \mathrm{LAC}(R^+), \qquad \lim_{t\to\infty} \operatorname{ess\,sup}_{s\ge t}|x'(s)|=0. \tag{1.32}$$

From the above one finally obtains an asymptotic result on the bounded solutions of (1.1) with μ assumed only locally finite and without (1.11), (1.12).

THEOREM 4. *Assume* (1.3) *and* (1.19) *hold. Let $r\in \mathrm{LAC}(R^+)$ satisfy* (1.9), (1.10) *and suppose*

$$\lim_{t\to\infty} r(t)=\lim_{t\to\infty} (r*f)(t)=0. \tag{1.33}$$

Also let

$$\mathrm{Im}\,\hat{\mu}(\omega)=0 \quad \text{for } \omega\in Z \overset{\text{def}}{=} \{\omega \mid \omega\neq 0,\ \mathrm{Re}\,\hat{\mu}(\omega)=0\}, \tag{1.34}$$

$$xg(x)>0, \qquad x\neq 0, \tag{1.35}$$

$$\liminf_{|x|\to 0} x^{-1}g(x)>0. \tag{1.36}$$

Finally suppose that $x\in(LAC\cap L^\infty)(R^+)$ satisfies (1.1). *Then*

$$\lim_{t\to\infty} x(t)=0. \tag{1.37}$$

If in addition

$$\lim_{t\to\infty} \operatorname{ess\,sup}_{s\ge t}|f(s)|=0, \tag{1.38}$$

then

$$\lim_{t\to\infty} \operatorname{ess\,sup}_{s\ge t}|x'(s)|=0. \tag{1.39}$$

As the solution z of (1.2) is given by $z=x_0 r+r*f$, it is clear that the assumption (1.33) implies $\lim_{t\to\infty} z(t)=0$. Analogously, (1.10), (1.38) yield $\lim_{t\to\infty} \operatorname{ess\,sup}_{s\ge t}|z'(s)|=0$. Concerning the size of r we note that $r\in L^1(R^+)$ is not explicitly required in Theorem 4, only (1.33).

Our last result concerns the existence of bounded solutions of (1.1).

THEOREM 5. *Assume* (1.3), (1.34) *hold and let*

$$|g(x)|\leq c[1+G(x)], \quad G(x)\geq \varepsilon x^2-c, \quad x\in R, \tag{1.40}$$

for some c, $\varepsilon>0$, *where* $G(x)\overset{\text{def}}{=}\int_0^x g(u)\,du$. *Let* x, r *be locally absolutely continuous solutions of* (1.1), (1.9) *respectively and suppose that*

$$r, r'\in L^1(R^+), \tag{1.41}$$

$$f\in \mathrm{LAC}(R^+), \qquad f'\in L^1(R^+). \tag{1.42}$$

Then

$$\sup_{t\in R^+}|x(t)|<\infty. \tag{1.43}$$

Earlier boundedness results on (1.1) with positive definite kernels (see [6], [11]) have required $f\in L^p(R^+)$, with $p=1$, 2. Obviously Theorem 5 allows much larger nonhomogeneous terms.

2. Proof of Theorem 1. Convolve (1.1) with r and use (1.9). This gives

$$r*x'-r'*g(x)=r*f. \tag{2.1}$$

Note that if both f_1, f_2 are measurable functions defined on R^+, then $f_1*f_2\overset{\text{def}}{=}\int_0^t f_1(t-s)f_2(s)\,ds$. An integration of the first term on the left side of (2.1) by parts results in

$$x(t)-\int_0^t h(x(t-s))r'(s)\,ds=z(t), \qquad t\in R^+, \tag{2.2}$$

where $h(x)\overset{\text{def}}{=}g(x)-x$, $x\in R$, and where we have used the fact that $z=x_0 r+r*f$. Differentiate (2.2) to obtain

$$x'(t)+\int_{[0,t]} h(x(t-s))\,d\nu(s)=z'(t) \quad \text{a.e. on } R^+. \tag{2.3}$$

From (1.9), (1.10) and from the definition of ν, it follows after straightforward computations ($\hat{\nu}=\int_{R^+}e^{-i\omega t}\,d\nu(t)$) that

$$\mathrm{Re}\,\hat{\nu}(\omega)=\omega^2\,\mathrm{Re}\,\hat{\mu}(\omega)|i\omega+\hat{\mu}(\omega)|^{-2}, \qquad \omega\neq 0, \tag{2.4}$$

$$\mathrm{Im}\,\hat{\nu}(\omega)=\omega^2\,\mathrm{Im}\,\hat{\mu}(\omega)|i\omega+\hat{\mu}(\omega)|^{-2}+\omega|\hat{\mu}(\omega)|^2|i\omega+\hat{\mu}(\omega)|^{-2}, \qquad \omega\neq 0, \tag{2.5}$$

$$\hat{\nu}(0)=0. \tag{2.6}$$

As μ is positive definite we have $\mathrm{Re}\,\hat{\mu}\geq 0$, $\omega\in R$, $\omega\neq 0$, and hence

$$\mathrm{Re}\,\hat{\nu}(\omega)\geq 0, \qquad \omega\in R, \tag{2.7}$$

and by (1.12), (2.4), (2.6)

$$\mathrm{Re}\,\hat{\nu}(\omega)=0 \quad \text{iff } \omega\in Z\cup\{0\}. \tag{2.8}$$

But by (1.13), (2.5), (2.6)

$$\mathrm{Im}\,\hat{\nu}(\omega)=0 \quad \text{if } \omega\in Z\cup\{0\}. \tag{2.9}$$

From (1.3), (1.10), (1.11), (1.14), (1.15), (2.7)–(2.9) it follows that we may apply [12, Cor. 3b] to the equation (2.2). This gives (1.16) and

$$\lim_{t\to\infty}\{x(t)+h(x(t))[1-r(\infty)]\}=z(\infty). \tag{2.10}$$

Substitute the expression for $h(x)$ to get (1.17). Provided $z'\in L^\infty(R^+)$, $\lim_{t\to\infty}\operatorname{ess\,sup}_{s\geq t}|z'(s)|=0$, we obtain $\lim_{t\to\infty}\operatorname{ess\,sup}_{s\geq t}|x'(s)|=0$ by applying [12, Thm. 1b] to (2.3).

3. Proof of Theorem 2. We begin by demonstrating that

$$\sup_{y\in Y_K}\|g(y(\tau))\|_{L^2(R)}<\infty. \tag{3.0}$$

This will occupy us until the beginning of the paragraph containing (3.47).

For $t>0$ we define

$$m_t^2=\sup_{y\in Y_K}\int_{-t}^{t}|g(y(\tau))|^2\,d\tau. \tag{3.1}$$

Assume $\lim_{t\to\infty}m_t^2=\infty$, otherwise (3.0) holds. Then choose for each $t>0$, $y_t\in Y_k$ such that

$$\int_{-t}^{t}|g(y_t(\tau))|^2\,d\tau=m_t^2. \tag{3.2}$$

As Y_K is translation invariant, one also has

$$\sup_{\substack{y\in Y_K\\ s\in R}}\int_{s-t}^{s+t}|g(y(\tau))|^2\,d\tau=\sup_{s\in R}\int_{s-t}^{s+t}|g(y_t(\tau))|^2\,d\tau=m_t^2. \tag{3.3}$$

Take $T>0$ (we will later choose T sufficiently large) and let $t>T$. In the estimates which follow we repeatedly obtain upper bounds f_i, which are functions of T. Each function $f_i(T)$ is a priori given by g, α and K. In particular note that each f_i is independent of t and y_t. An odd-indexed bound $f_{2n+1}(T)$ is always a monotonically decreasing function of T and satisfies

$$\lim_{T\to\infty}f_{2n+1}(T)=0,$$

whereas an even-indexed bound $f_{2n}(T)$ satisfies $f_{2n}\in L^\infty_{\mathrm{loc}}(R^+)$.

Multiply (1.26) by $g(y_t(\tau))$, integrate over $[-t,t]$, split the integral term into two parts and define z_t, g_K, G_K by

$$z_t(\tau)=g(y_t(\tau)),\quad |\tau|\leq t,\qquad z_t(\tau)=0,\quad |\tau|>t, \tag{3.4}$$

$g_K=\sup_{|x|\leq K}|g(x)|$, $G_K=\sup_{|x|\leq K}|G(x)|$. This gives, after an application of Parseval's relation,

$$(2\pi)^{-1}\int_R|\hat z_t|^2\operatorname{Re}\hat\alpha(\omega)\,d\omega\leq 2G_K+\left|\int_{-t}^{t}g(y_t(\tau))\int_{(\tau+t,\infty)}g(y_t(\tau-s))\,d\alpha(s)\,d\tau\right|. \tag{3.5}$$

As α is positive definite, one has by (3.3) and after estimating the right side of (3.5) (see [4, Assertion 1])

$$\int_R|\hat z_t\operatorname{Re}\hat\alpha|^2\,d\omega\leq m_t^2f_1(T)+f_2(T),\qquad t>T. \tag{3.6}$$

Define u_t, f_t by

$$u_t(\tau)=y_t'(\tau), \quad |\tau|\le t, \qquad u_t(\tau)=0, \quad |\tau|>t, \tag{3.7}$$

$$f_t(\tau)=\begin{cases} 0, & \tau<-t, \\ -\int_{(\tau+t,\infty)} g(y_t(\tau-s))\,d\alpha(s), & |\tau|\le t, \\ \int_{[\tau-t,\tau+t]} g(y_t(\tau-s))\,d\alpha(s), & \tau>t. \end{cases} \tag{3.8}$$

Then

$$u_t(\tau)+\int_R z_t(\tau-s)\,d\alpha(s)=f_t(\tau) \quad \text{a.e. on } R. \tag{3.9}$$

Note that as α is finite and u_t, z_t have compact support, then u_t, z_t, $f_t\in(L^1\cap L^2)(R)$, and so the Fourier transforms to follow are well defined.

Choose $\omega_0\in(0,1)$ such that (recall the second part of (1.23))

$$2\operatorname{Re}\hat a(\omega)\ge\hat a(0), \qquad |\omega|\le\omega_0, \tag{3.10}$$

and let $\lambda>0$ satisfy

$$|g(x)-g(y)|\le\lambda|x-y| \quad \text{for } |x|, |y|\le K. \tag{3.11}$$

Denote $\alpha_0 \overset{\text{def}}{=} \max(1, \sup_{\omega\in R}|\hat\alpha(\omega)|^2)$, $\beta \overset{\text{def}}{=} \inf_{\omega\in Z_2}\operatorname{Re}\hat a(\omega)$. Then take any $\varepsilon\in(0,1)$ such that

$$\varepsilon\le 8^{-1}\left[\lambda^2\omega_0^{-2}+\lambda^2\alpha_0\right]^{-1}, \tag{3.12}$$

$$2\operatorname{Re}\hat a(\omega)\ge\beta>0 \quad \text{for } \omega\in S_0 \quad \text{where} \tag{3.13}$$

$$S_0 \overset{\text{def}}{=} \left\{\omega\,|\,\operatorname{dist}(\omega,Z_2)\le\varepsilon,\ |\operatorname{Im}\hat\alpha|^2\ge\varepsilon,\ \omega_0<|\omega|\le\varepsilon^{-1}\right\}. \tag{3.14}$$

Divide R in four pairwise disjoint parts S_i as follows:

$$S_1 \overset{\text{def}}{=} \left\{\omega\,|\,|\omega|>\varepsilon^{-1}\right\}, \tag{3.15}$$

$$S_2 \overset{\text{def}}{=} \left\{\omega\,|\,\omega_0<|\omega|\le\varepsilon^{-1},\ |\operatorname{Im}\hat\alpha|^2<\varepsilon\right\}, \tag{3.16}$$

$$S_3 \overset{\text{def}}{=} \left\{\omega\,|\,\omega_0<|\omega|\le\varepsilon^{-1},\ |\operatorname{Im}\hat\alpha|^2\ge\varepsilon,\ \operatorname{dist}(\omega,Z_2)>\varepsilon\right\}, \tag{3.17}$$

$$S_4 \overset{\text{def}}{=} S_0\cup\left\{\omega\,|\,|\omega|\le\omega_0\right\}. \tag{3.18}$$

Note that $R=\cup_{i=1}^4 S_i$. In what follows $K_i(\varepsilon,T)$ will denote bounds which are independent of t and y_t, but which do depend on ε and T.

Our next goal is to show that there exists a constant c_1 (depending only on ω_0, λ, α_0 and in particular independent of t, y_t, ε, T) such that, provided T is fixed sufficiently large, then

$$\int_{R\setminus S_4}|\hat u_t|^2\,d\omega\le\varepsilon\int_{|\omega|\le\omega_0}|\hat z_t|^2\,d\omega+\varepsilon c_1\int_{S_4}|\hat u_t|^2\,d\omega+K_1(\varepsilon,T) \tag{3.19}$$

for $t>T$.

By (3.9)

$$\hat{u}_t(\omega)+\hat{z}_t(\omega)\hat{\alpha}(\omega)=\hat{f}_t(\omega), \qquad \omega\in R, \tag{3.20}$$

and so

$$2^{-1}|\hat{u}_t|^2\le|\hat{z}_t\hat{\alpha}|^2+|\hat{f}_t|^2. \tag{3.21}$$

Integrate (3.21) over $R\setminus S_4$ and estimate the right side. Obviously

$$\int_{S_1}|\hat{z}_t\hat{\alpha}|^2\,d\omega\le\alpha_0\int_{\varepsilon^{-1}\le|\omega|}|\hat{z}_t|^2\,d\omega, \tag{3.22}$$

$$\int_{S_2}|\hat{z}_t\hat{\alpha}|^2\,d\omega\le\varepsilon\int_{\omega_0\le|\omega|}|\hat{z}_t|^2\,d\omega+\int_R|\hat{z}_t\operatorname{Re}\hat{\alpha}|^2\,d\omega. \tag{3.23}$$

Then note that by (1.22), (1.23), (3.17) there exists $\delta=\delta(\varepsilon)\in(0,1)$ such that $\operatorname{Re}\hat{\alpha}(\omega)\ge\delta^{1/2}$, $\omega\in S_3$. Take any such δ. Then

$$\int_{S_3}|\hat{z}_t\hat{\alpha}|^2\,d\omega\le\alpha_0\delta^{-1}\int_R|\hat{z}_t\operatorname{Re}\hat{\alpha}|^2\,d\omega. \tag{3.24}$$

By slight modifications of the estimates of [4, Assertion 2] one gets

$$\int_R|\hat{f}_t|^2\,d\omega\le m_t^2f_3(T)+f_4(T), \qquad t>T. \tag{3.25}$$

From (3.21)–(3.25) and from (3.6) follows

$$\begin{aligned}2^{-1}\int_{R\setminus S_4}|\hat{u}_t|^2\,d\omega\le{}&\delta^{-1}m_t^2f_5(T)+\delta^{-1}f_6(T)\\&+\alpha_0\int_{\varepsilon^{-1}\le|\omega|}|\hat{z}_t|^2\,d\omega+\varepsilon\int_{\omega_0\le|\omega|}|\hat{z}_t|^2\,d\omega.\end{aligned} \tag{3.26}$$

Choose $g_n\in C^1(R)$ such that

$$|g_n'(x)|\le\lambda, \quad |x|\le K, \qquad \lim_{n\to\infty}\sup_{|x|\le K}|g_n(x)-g(x)|=0,$$

define α_{nt}, β_{nt} by

$$\alpha_{nt}(\tau)=\begin{cases}g_n(y_t(\tau)), & |\tau|\le t,\\ 0, & |\tau|>t,\end{cases}$$

$$\beta_{nt}(\tau)=\begin{cases}\dfrac{d}{d\tau}\left[g_n(y_t(\tau))\right], & |\tau|\le t,\\ 0, & |\tau|>t,\end{cases}$$

and observe that the three relations

$$\lim_{n\to\infty}\int_R|\hat{z}_t(\omega)-\hat{\alpha}_{nt}(\omega)|^2\,d\omega=0,$$

$$|\hat{\alpha}_{nt}(\omega)|^2\le4g_K^2|\omega|^{-2}+2|\omega|^{-2}|\hat{\beta}_{nt}(\omega)|^2, \qquad \omega\ne0,$$

$$|\beta_{nt}(\tau)|\le\lambda|u_t(\tau)|, \qquad \tau\in R,$$

are an easy consequence. Now use these three relations to estimate $\int_{0<\gamma\le|\omega|}|\hat{z}_t|^2\,d\omega$ upwards. Use the first to replace $\hat{z}_t$ by $\hat{\alpha}_{nt}$, the second for the step from $\hat{\alpha}_{nt}$ to $\omega^{-1}\hat{\beta}_{nt}$. Finally, by the third,

$$\int_{0<\gamma\le|\omega|}\left|\frac{\hat{\beta}_{nt}}{\omega}\right|^2 d\omega\le\gamma^{-2}\int_R|\hat{\beta}_{nt}|^2\,d\omega=2\pi\gamma^{-2}\int_R|\beta_{nt}|^2\,d\tau$$

$$\le 2\pi\lambda^2\gamma^{-2}\int_R|u_t|^2\,d\tau=\lambda^2\gamma^{-2}\int_R|\hat{u}_t|^2\,d\omega.$$

Hence, for any $\gamma>0$,

$$\int_{\gamma\le|\omega|}|\hat{z}_t|^2\,d\omega\le 2\lambda^2\gamma^{-2}\int_R|\hat{u}_t|^2\,d\omega+4g_K^2\gamma^{-1}. \tag{3.27}$$

Use (3.27) (with $\gamma=\omega_0$, ε^{-1}) to estimate the right side of (3.26). (Note that

$$m_t^2=\int_{\omega_0<|\omega|}|\hat{z}_t|^2\,d\omega+\int_{|\omega|\le\omega_0}|\hat{z}_t|^2\,d\omega.\ )$$

This yields

$$4^{-1}\int_{R\setminus S_4}|\hat{u}_t|^2\,d\omega\le\delta^{-1}f_5(T)\int_{|\omega|\le\omega_0}|\hat{z}_t|^2\,d\omega+\delta^{-1}f_7(T)\int_R|\hat{u}_t|^2\,d\omega \tag{3.28}$$

$$+\varepsilon c_0\int_{S_4}|\hat{u}_t|^2\,d\omega+\delta^{-1}f_8(T)$$

where we have also used (3.12) and defined $c_0=2\alpha_0\lambda^2+2\lambda^2\omega_0^{-2}$. Choose T sufficiently large so that

$$\delta^{-1}f_7(T)\le 8^{-1}\varepsilon,\qquad \delta^{-1}f_5(T)\le 8^{-1}\varepsilon.$$

From (3.28) one then has, for $t>T$ (recall that $\varepsilon<1$),

$$\int_{R\setminus S_4}|\hat{u}_t|^2\,d\omega\le\varepsilon\int_{|\omega|\le\omega_0}|\hat{z}_t|^2\,d\omega+\varepsilon[1+8c_0]\int_{S_4}|\hat{u}_t|^2\,d\omega+8\delta^{-1}f_8(T), \tag{3.29}$$

and so (3.19) holds, with $c_1=1+8c_0$ and $K_1=8\delta^{-1}f_8$.

In what follows we wish to eliminate the second integral on the right side of (3.29). Thus we show that there exists a constant c_2 (depending only on ω_0, λ, α_0) such that provided T is fixed sufficiently large then

$$\int_{R\setminus S_4}|\hat{u}_t|^2\,d\omega\le\varepsilon c_2\int_{S_4}|\hat{z}_t|^2\,d\omega+K_2(\varepsilon,T),\qquad t>T. \tag{3.30}$$

By (3.21), (3.25), provided T is taken so that $f_3(T)\le\alpha_0$,

$$2^{-1}\int_{S_4}|\hat{u}_t|^2\,d\omega\le\int_{S_4}|\hat{z}_t\hat{\alpha}|^2\,d\omega+\int_{S_4}|\hat{f}_t|^2\,d\omega \tag{3.31}$$

$$\le 2\alpha_0\int_{S_4}|\hat{z}_t|^2\,d\omega+f_3(T)\int_{R\setminus S_4}|\hat{z}_t|^2\,d\omega+f_4(T).$$

Invoke (3.27) with $\gamma=\omega_0$ and then (3.19) to obtain

$$\int_{R\setminus S_4}|\hat{z}_t|^2\,d\omega\le\int_{\omega_0\le|\omega|}|\hat{z}_t|^2\,d\omega \tag{3.32}$$
$$\le 2\lambda^2\omega_0^{-2}\varepsilon\int_{|\omega|\le\omega_0}|\hat{z}_t|^2\,d\omega+\tilde{c}_1\int_{S_4}|\hat{u}_t|^2\,d\omega+\tilde{K}_2(\varepsilon,T)$$

where $\tilde{c}_1=2\lambda^2\omega_0^{-2}[1+c_1]$; $\tilde{K}_2=2\lambda^2\omega_0^{-2}K_1+4g_K^2\omega_0^{-1}$. Now use (3.32) to estimate the last integral on the right side of (3.31). This yields

$$\int_{S_4}|\hat{u}_t|^2\,d\omega\le 12\alpha_0\int_{S_4}|\hat{z}_t|^2\,d\omega+4f_4(T)+4f_3(T)\tilde{K}_2(\varepsilon,T),\qquad t>T, \tag{3.33}$$

provided T is taken such that $f_3(T)\tilde{c}_1\le 4^{-1}$, $f_3(T)2\lambda^2\omega_0^{-2}\varepsilon\le\alpha_0$. Finally estimate the right side of (3.19) with the aid of (3.33). The relation (3.30) follows, with $c_2=1+12\alpha_0c_1$.

Take $\gamma=\omega_0$ in (3.27), add $\int_{|\omega|\le\omega_0}|\hat{z}_t|^2\,d\omega$ to both sides and use (3.30), (3.33) to estimate the right side. One obtains

$$\int_R|\hat{z}_t|^2\,d\omega\le\tilde{c}_2\int_{S_4}|\hat{z}_t|^2\,d\omega+K_3(\varepsilon,T),\qquad t>T, \tag{3.34}$$

where $\tilde{c}_2=1+2\lambda^2\omega_0^{-2}[c_2+12\alpha_0]$. Use (3.34) in (3.25) to get

$$\int_R|\hat{f}_t|^2\,d\omega\le\tilde{c}_2f_3(T)\int_{S_4}|\hat{z}_t|^2\,d\omega+K_4(\varepsilon,T),\qquad t>T, \tag{3.35}$$

where $K_4=f_3K_3+f_4$.

By (3.20) $|\hat{z}_t\hat{\alpha}|^2\le 2|\hat{u}_t|^2+2|\hat{f}_t|^2$. Integrate this inequality over $R\setminus S_4$ and invoke (3.30), (3.35). This yields

$$\int_{R\setminus S_4}|\hat{z}_t\hat{\alpha}|^2\,d\omega\le 2\varepsilon[c_2+\tilde{c}_2]\int_{S_4}|\hat{z}_t|^2\,d\omega+K_5(\varepsilon,T),\qquad t>T, \tag{3.36}$$

provided T is taken such that $f_3(T)\le\varepsilon$. But $|\hat{\alpha}|=|\omega\hat{a}|$ and hence

$$\int_{R\setminus S_4}|\hat{z}_t\operatorname{Re}\hat{a}|^2\,d\omega\le\omega_0^{-2}\int_{R\setminus S_4}|\hat{z}_t\hat{\alpha}|^2\,d\omega, \tag{3.37}$$

which together with (3.36) implies

$$\int_{R\setminus S_4}|\hat{z}_t\operatorname{Re}\hat{a}|^2\,d\omega\le\varepsilon c_3^2\int_{S_4}|\hat{z}_t|^2\,d\omega+\omega_0^{-2}K_5(\varepsilon,T), \tag{3.38}$$

for $t>T$ and where $c_3^2=2\omega_0^{-2}[c_2+\tilde{c}_2]$.

Define A_ε, B_ε by

$$A_\varepsilon=\{\omega|\omega\in R\setminus S_4,\ |\operatorname{Re}\hat{a}(\omega)|\ge c_3\varepsilon^{1/2}\}, \tag{3.39}$$

$$B_\varepsilon=\{\omega|\omega\in R\setminus S_4,\ |\operatorname{Re}\hat{a}(\omega)|<c_3\varepsilon^{1/2}\}. \tag{3.40}$$

A combination of (3.38), (3.39) results in

$$\left|\int_{A_\varepsilon}\operatorname{Re}\hat{a}|\hat{z}_t|^2\,d\omega\right|\le c_3^{-1}\varepsilon^{-1/2}\int_{A_\varepsilon}|\hat{z}_t\operatorname{Re}\hat{a}|^2\,d\omega\le c_3\varepsilon^{1/2}\int_{S_4}|\hat{z}_t|^2\,d\omega+K_6(\varepsilon,T), \tag{3.41}$$

and from (3.34), (3.40) follows

$$\left|\int_{B_\varepsilon} \operatorname{Re}\hat{a}|\hat{z}_t|^2\,d\omega\right| \le c_3\varepsilon^{1/2}\int_{R\setminus S_4}|\hat{z}_t|^2\,d\omega \le c_3\tilde{c}_2\varepsilon^{1/2}\int_{S_4}|\hat{z}_t|^2\,d\omega + K_7(\varepsilon,T). \tag{3.42}$$

Consequently

$$\left|\int_{R\setminus S_4} \operatorname{Re}\hat{a}|\hat{z}_t|^2\,d\omega\right| \le c_4\varepsilon^{1/2}\int_{S_4}|\hat{z}_t|^2\,d\omega + K_8(\varepsilon,T), \qquad t>T, \tag{3.43}$$

where $c_4 = c_3(1+\tilde{c}_2)$.

Multiply (1.18) by z_t, integrate over $[-t,t]$ and use Parseval's relation. This gives

$$\int_{-t}^{t} z_t(\tau)y_t(\tau)\,d\tau + \int_R |\hat{z}_t(\omega)|^2 \operatorname{Re}\hat{a}(\omega)\,d\omega = -\int_{-t}^{t} z_t(\tau)\int_{(\tau+t,\infty)} g(y(\tau-s))a(s)\,ds\,d\tau. \tag{3.44}$$

The right side of (3.44) ($\overset{\text{def}}{=} r(t)$) can be shown (use [4, Assertion 1] as in (3.5), (3.6) together with (3.34)) to satisfy

$$|r(t)| \le f_5(T)\int_{S_4}|\hat{z}_t|^2\,d\omega + K_9(\varepsilon,T), \qquad t>T. \tag{3.45}$$

Combine (3.43)–(3.45), use $yg(y)\ge 0$, $|y|\le K$ (note that this is the first place where this condition is used) and recall that by (3.10), (3.13), (3.18) $2\operatorname{Re}\hat{a}(\omega)\ge\beta>0$, $\omega\in S_4$. This yields

$$\frac{\beta}{2}\int_{S_4}|\hat{z}_t|^2\,d\omega \le \varepsilon^{1/2}[c_4+1]\int_{S_4}|\hat{z}_t|^2\,d\omega + K_{10}(\varepsilon,T), \qquad t>T, \tag{3.46}$$

provided T is taken such that $f_5(T)\le\varepsilon^{1/2}$. But $\lim_{t\to\infty} m_t^2=\infty$ and so, by (3.34), we have $\lim_{t\to\infty}\int_{S_4}|\hat{z}_t|^2\,d\omega=\infty$. An examination of (3.46) reveals that this implies $\beta\le 2\varepsilon^{1/2}[c_4+1]$. The constants c_4 and β are, however, independent of ε, which was taken sufficiently small but otherwise arbitrary, and hence a contradiction follows. We conclude that $\sup_{t>0} m_t^2<\infty$.

By (3.0) and (1.18), (1.21) we have

$$\sup_{y\in Y_K}\|y\|_{L^2(R)}<\infty. \tag{3.47}$$

From (1.20), (1.26) and as $g(y(t))\in L^\infty(R)$,

$$\operatorname*{ess\,sup}_{s\in R}|y'(s)|<\infty, \qquad y\in Y_K. \tag{3.48}$$

A combination of (3.47), (3.48) implies $y\in(L^2\cap \mathrm{BUC})(R)$ and so

$$y(\infty)=y(-\infty)=0, \qquad y\in Y_K. \tag{3.49}$$

Multiply (1.26) by $g(y(t))$, integrate and use Parseval's relation. This gives—by (1.20), (1.21), (3.0), (3.47) all the integrals below are well defined—

$$G(y(\infty))-G(y(-\infty))+(2\pi)^{-1}\int_R|\hat{g}|^2\operatorname{Re}\hat{\alpha}\,d\omega=0, \tag{3.50}$$

which by (3.49) yields

$$\int_R |\hat{g}|^2 \operatorname{Re}\hat{\alpha}\, d\omega = 0, \qquad y \in Y_K. \tag{3.51}$$

Suppose (1.25) does not hold and let $y \in Y_K$, $y(t) \not\equiv 0$. Then $y'(t) \not\equiv 0$, and so by (1.26) $\int_{R^+} g(y(t-s))\, d\alpha(s) \not\equiv 0$, which shows that $g(y(t)) \not\equiv 0$. Consequently

$$\int_R |\hat{g}(\omega)|^2\, d\omega > 0. \tag{3.52}$$

Define $S = \{\omega | \hat{g}(\omega) \neq 0\}$. From (3.51) and the definition of Z we have $m(\{\omega | \omega \in S, \omega \notin Z\}) = 0$. But $Z = Z_1 \cup Z_2 \cup \{0\}$ and by (1.22) $m(Z_1 \cup \{0\}) = 0$. Hence

$$m(\{\omega | \omega \in S, \omega \notin Z_2\}) = 0, \tag{3.53}$$

and

$$\int_R |\hat{g}|^2 \operatorname{Re}\hat{a}\, d\omega = \int_S |\hat{g}|^2 \operatorname{Re}\hat{a}\, d\omega = \int_{Z_2} |\hat{g}|^2 \operatorname{Re}\hat{a}\, d\omega \geq \beta \int_{Z_2} |\hat{g}|^2\, d\omega \tag{3.54}$$

where $\beta = \inf_{\omega \in Z_2} \operatorname{Re}\hat{a}(\omega) > 0$. Multiply (1.18)by $g(y(t))$, integrate and use Parseval's relation to obtain the first equality in (3.55). The first inequality follows by (1.24) and the second by (3.54). Finally observe that the second equality is a consequence of (3.53).

$$\begin{aligned} 0 \leq \int_R y(\tau) g(y(\tau))\, d\tau &= -(2\pi)^{-1} \int_R |\hat{g}|^2 \operatorname{Re}\hat{a}\, d\omega \\ &\leq -\frac{\beta}{2\pi} \int_{Z_2} |\hat{g}|^2\, d\omega = -\frac{\beta}{2\pi} \int_R |\hat{g}|^2\, d\omega. \end{aligned} \tag{3.55}$$

Thus $\int_R |\hat{g}|^2\, d\omega = 0$. This, however, violates (3.52), and consequently our assumption $y \not\equiv 0$ is false and (1.25) holds.

4. Proof of Theorem 4. We begin by proving two auxiliary lemmas.

LEMMA 2. *Let μ satisfy* (1.3ii) *and let, for $\lambda > 0$, $r_\lambda \in \mathrm{LAC}(R^+)$ be the solution of*

$$r_\lambda'(t) + \lambda (r_\lambda * \mu)(t) = 0 \quad a.e.\ on\ R^+, \quad r_\lambda(0) = 1, \tag{4.1}$$

and define $r(t) = r_1(t)$. Then if

$$r' \in L^1(R^+) \tag{4.2}$$

and if (1.34) *holds, one has*

$$r_\lambda' \in L^1(R^+) \quad for\ \lambda > 0. \tag{4.3}$$

If in addition

$$r' \in \mathrm{NBV}(R^+), \tag{4.4}$$

then

$$r_\lambda' \in \mathrm{NBV}(R^+) \quad for\ \lambda > 0. \tag{4.5}$$

Proof of Lemma 2. Multiply (4.1) by r_λ and integrate over $[0, t]$. The result is

$$r_\lambda^2(t) - 1 + 2\lambda \int_0^t r_\lambda (r_\lambda * \mu)(\tau)\, d\tau = 0. \tag{4.6}$$

As μ is a positive definite measure we conclude from (4.6) that $|r_\lambda(t)|\le 1$, $t\in R^+$, $\lambda>0$. Thus $\tilde r_\lambda(s)\overset{\text{def}}{=}\int_{R^+}e^{-st}r_\lambda(t)\,dt$ is well defined for $\operatorname{Re}s>0$. But then, as $\tilde\mu(s)$ exists for $\operatorname{Re}s>0$, we have that $\tilde r_\lambda'(s)$ is well defined for $\operatorname{Re}s>0$ and satisfies

$$-\tilde r_\lambda'(s)=\frac{\lambda\tilde\mu(s)}{s+\lambda\tilde\mu(s)},\qquad \operatorname{Re}s>0,\quad \lambda>0. \tag{4.7}$$

($\operatorname{Re}\tilde\mu(s)\ge 0$ implies $s+\lambda\tilde\mu(s)\ne 0$ for $\operatorname{Re}s>0$). As in §1 note that because by (4.2) $\tilde\mu(s)/(s+\tilde\mu(s))$ is the transform of $h\overset{\text{def}}{=}-r'\in L^1(R^+)$, then

$$\tilde h(s)\overset{\text{def}}{=}\begin{cases}\dfrac{\tilde\mu(s)}{s+\tilde\mu(s)}, & \operatorname{Re}s>0,\\[2ex] \lim\limits_{\substack{z\to s\\ \operatorname{Re}z>0}}\dfrac{\tilde\mu(z)}{z+\tilde\mu(z)}, & \operatorname{Re}s=0,\end{cases}$$

is well defined and continuous for $\operatorname{Re}s\ge 0$. Consequently $\hat\mu(\omega)\overset{\text{def}}{=}\lim_{s\to i\omega,\operatorname{Re}s>0}\tilde\mu(s)$ exists, possibly infinite, for $\omega\in R$, $\omega\ne 0$. Clearly $\hat h(0)=-\int_{R^+}r'(t)\,dt=1-r(\infty)$. We claim that

$$0\le\hat h(0)\le 1, \tag{4.8}$$

or equivalently $0\le r(\infty)\le 1$. The fact that $\hat h(0)\ge 0$ ($r(\infty)\le 1$) was already established. Suppose $\hat h(0)>1$. Then by the continuity of $\tilde h$ there exists $x>0$ such that $\operatorname{Re}\tilde\mu(x)/(x+\tilde\mu(x))>1$. But this implies $x+\operatorname{Re}\tilde\mu(x)<0$ which obviously cannot hold. Thus (4.8) is satisfied.

Next observe that

$$\frac{\lambda\tilde\mu(s)}{s+\lambda\tilde\mu(s)}=\frac{\lambda\tilde h(s)}{1+(\lambda-1)\tilde h(s)},\qquad \operatorname{Re}s>0. \tag{4.9}$$

As $\operatorname{Re}\tilde\mu(s)\ge 0$ for $\operatorname{Re}s>0$, one immediately has $1+(\lambda-1)\tilde h(s)\ne 0$ for $\operatorname{Re}s>0$. Suppose $1+(\lambda-1)\tilde h(i\omega)=0$ for some $\omega=\omega_0\ne 0$. Clearly this cannot hold if $|\hat\mu(\omega_0)|=\infty$. Thus let $|\hat\mu(\omega_0)|<\infty$. But then $i\omega_0+\lambda\hat\mu(\omega_0)=0$ which by (1.34) and as μ is positive definite is excluded. Therefore, recalling also (4.8),

$$1+(\lambda-1)\tilde h(s)\ne 0,\qquad \operatorname{Re}s\ge 0. \tag{4.10}$$

As $h\in L^1(R^+)$ we now have by (4.7), (4.9), (4.10) and an application of a result of Paley–Wiener [7] that (4.3) holds.

To prove (4.5) we note that easy calculations give

$$r_\lambda'(t)=(\lambda-1)(r'*r_\lambda')(t)+\lambda r'(t). \tag{4.11}$$

But by (4.3), (4.4) $(r'*r_\lambda')\in\mathrm{NBV}(R^+)$ and so from (4.4), (4.11) we have (4.5).

LEMMA 3. *Let μ, f satisfy* (1.3ii), (1.3iii), *let r_λ, r be as in Lemma* 2 *and assume* (1.34), (4.2) *hold. Then if*

$$\lim_{t\to\infty}r(t)=\lim_{t\to\infty}(r*f)(t)=0 \tag{4.12}$$

one has

$$\lim_{t\to\infty}r_\lambda(t)=\lim_{t\to\infty}(r_\lambda*f)(t)=0,\qquad \lambda>0. \tag{4.13}$$

Proof of Lemma 3. Straightforward calculations show that

$$r_\lambda(t)=\left(\frac{\lambda-1}{\lambda}\right)(r'_\lambda * r)(t)+r(t). \tag{4.14}$$

Hence the first part of (4.13) follows from (4.3) and from the first part of (4.12). To get the second part of (4.13) it suffices to convolve (4.14) by f and to apply (4.3) and the second part of (4.12).

Proof of Theorem 4. By the above lemmas and by (1.3), (1.10), (1.33), (1.34) we have

$$r'_\lambda \in (L^1 \cap \mathrm{NBV})(R^+), \qquad \lambda>0, \tag{4.15}$$

$$z_\lambda(t)=0, \qquad \lambda>0, \tag{4.16}$$

where $z_\lambda = x_0 r_\lambda + r_\lambda * f$.

Convolve (1.1) with r_λ and use (4.1). Perform an integration by parts and define $h_\lambda(x)=\lambda^{-1}g(x)-x$, $x\in R$. Let ν_λ be the measure corresponding to $-r'_\lambda$; thus $\nu_\lambda([0,t]) = -r'_\lambda(t)$, $t\geq 0$. This gives

$$x(t)+\int_{[0,t]} h_\lambda(x(t-s))\nu_\lambda([0,s])\,ds = z_\lambda(t), \qquad t\geq 0. \tag{4.17}$$

We wish to apply Theorem 3 to (4.17). From (1.19) follows

$$h_\lambda(x) \text{ is locally Lipschitzian}, \quad x\in R, \quad \lambda>0, \tag{4.18}$$

and invoking (1.35), (1.36) one has

$$xh_\lambda(x)\geq 0 \quad \text{for } |x|\leq \sup_{t\in R^+}|x(t)|, \tag{4.19}$$

provided λ is taken sufficiently small.

The relations (2.4)–(2.6) hold with μ replaced by $\lambda\mu$, ν with ν_λ and with the usual interpretation if $|\hat{\mu}(\omega)|=\infty$ for $\omega\neq 0$. Hence ($\hat{\nu}_\lambda \overset{\text{def}}{=} \int_{R^+} e^{-i\omega t}\,d\nu_\lambda(t)$)

$$\mathrm{Re}\,\hat{\nu}_\lambda(\omega)\geq 0, \qquad \omega\in R, \quad \lambda>0, \tag{4.20}$$

with

$$\mathrm{Re}\,\hat{\nu}_\lambda(\omega)=0 \quad \text{iff } \omega\in Z_1\cup Z_2\cup\{0\}, \tag{4.21}$$

where

$$Z_1 \overset{\text{def}}{=} \{\omega \mid \omega\neq 0,\ \mathrm{Re}\,\hat{\mu}(\omega)=0,\ |\hat{\mu}(\omega)|<\infty\}, \tag{4.22}$$

$$Z_2 \overset{\text{def}}{=} \{\omega \mid \omega\neq 0,\ |\hat{\mu}(\omega)|=\infty\}.$$

By (1.34)

$$\mathrm{Im}\,\hat{\nu}_\lambda(\omega)=0, \qquad \omega\in Z_1, \tag{4.23}$$

and using (4.3) and the first part of (4.13)

$$\mathrm{Im}\,\hat{\nu}_\lambda(\omega)=\omega, \qquad \omega\in Z_2, \tag{4.24}$$

$$\int_{R^+} e^{-i\omega t}\nu_\lambda([0,t])\,dt=1, \qquad \omega\in Z_2\cup\{0\}.$$

By (4.15), (4.16), (4.18)–(4.24), an application of the first part of Theorem 3 to (4.17) is permitted. This gives $\lim_{t\to\infty} x(t)=0$ and so also $\lim_{t\to\infty} g(x(t))=0$.

To complete the proof we note that

$$z_\lambda'(t)=x_0 r_\lambda'(t)+f(t)+(r_\lambda'*f)(t) \quad \text{a.e. on } R^+, \quad \lambda>0. \tag{4.25}$$

By (1.38), (4.15), (4.25) we have $\lim_{t\to\infty} \operatorname{ess\,sup}_{s\geq t}|z_\lambda'(s)|=0$. Thus an application of the second part of Theorem 3 is allowed and (1.39) follows.

5. Proof of Theorem 5. The method of the previous section enables us to transform (1.1) into

$$x'(t)+\int_{[0,t]} h_\lambda(x(t-s))\,d\nu_\lambda(s)=z_\lambda'(t) \quad \text{a.e. on } R^+, \tag{5.1}$$

where h_λ, ν_λ, z_λ are as in the proof of Theorem 4. From Lemma 2—which by (1.3), (1.41) can be applied—we have

$$r_\lambda'\in L^1(R^+), \qquad \lambda>0. \tag{5.2}$$

By (1.41), (4.14), (5.2)

$$r_\lambda\in L^1(R^+), \qquad \lambda>0. \tag{5.3}$$

From (1.42), (5.2), (5.3) follows

$$z_\lambda'\in L^1(R^+), \qquad \lambda>0. \tag{5.4}$$

Multiply (5.1) by $h_\lambda(x(t))$, integrate with respect to t over $[0,T]$ and use the fact that ν_λ is a positive definite measure. This yields

$$H_\lambda(x(T))-H_\lambda(x(0))\leq\int_0^T h_\lambda(x(t))z_\lambda'(t)\,dt, \tag{5.5}$$

where $H_\lambda(x)=\lambda^{-1}G(x)-2^{-1}x^2$. Making use of (1.40) one shows that for any sufficiently small λ there exist constants c_1, c_2 (depending on λ) such that $|h_\lambda(x)|\leq c_1+c_2 H_\lambda(x)$, $x\in R$. Therefore, by a simple application of Gronwall's inequality to (5.5) and recalling (5.4) we get

$$\sup_{T>0} H_\lambda(x(T))<\infty. \tag{5.6}$$

But (5.6), the second part of (1.40) and the definition of H_λ imply

$$\sup_{t\in R^+} |x(t)|<\infty. \tag{5.7}$$

We finally point out that under the present assumptions one also has, for all sufficiently small λ,

$$\sup_{T>0}\int_0^T h_\lambda(\tau)(h_\lambda*\nu_\lambda)(\tau)\,d\tau<\infty,$$

from which various consequences concerning the asymptotic behavior of $x(t)$ can be deduced.

REFERENCES

[1] R. W. Carr and K. B. Hannsgen, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961–983.

[2] G. Gripenberg, *On nonlinear Volterra equations with nonintegrable kernels*, this Journal, 11 (1980), pp. 668–682.

[3] S.-O. Londen, *On a Volterra integrodifferential equation with L^∞-perturbation and noncountable zero-set of the transformed kernel*, J. Integral Equations, 1 (1979), pp. 275–280.

[4] _____, *On an integral equation with L^∞-perturbation*, J. Integral Equations, to appear.
[5] _____, *Asymptotic properties of Volterra equations with nonintegrable kernels*, MRC Tech. Sum. Rep. 2152, Univ. of Wisconsin Mathematics Research Center, Madison, WI, 1980.
[6] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Adv. in Math., 22 (1976), pp. 278–304.
[7] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, RI, 1934.
[8] 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.
[9] O. J. STAFFANS, *Positive definite measures with applications to a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 219–237.
[10] _____, *Tauberian theorems for a positive definite form, with applications to a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 239–259.
[11] _____, *Boundedness and asymptotic behavior of solutions of a Volterra equation*, Michigan Math. J., 24 (1977), pp. 77–95.
[12] _____, *On a nonlinear integral equation with a nonintegrable perturbation*, J. Integral Equations, 1 (1979), pp. 291–307.
[13] D. V. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 14, No. 6, November 1983

© 1983 Society for Industrial and Applied Mathematics
0036-1410/83/1406/0012 $02.25/0

UNIFORM ASYMPTOTIC EXPANSIONS OF A CLASS OF MEIJER *G*-FUNCTIONS FOR A LARGE PARAMETER*

JERRY L. FIELDS†

Abstract. Asymptotic estimates of certain Meijer *G*-functions are derived using contour integration techniques. Making use of these results, asymptotic estimates of the generalized Jacobi functions are derived.

Notation and introduction. Throughout this work, extensive use will be made of the following conventions and hypergeometric notation.

As usual, the logarithm of a real positive number is a real number, unless stated otherwise, and if the functions $f(z)$, $g(z)$ are well defined, complex-valued, analytic functions in a simply connected domain $\mathfrak{D}$, possibly on a Riemann surface, where $f(z)\neq 0$, then the multiple-valued function

$$\{f(z)\}^{g(z)}=\exp\{g(z)\log f(z)\}=\exp\{g(z)[\log|f(z)|+i\arg f(z)]\},$$

can be defined uniquely in $\mathfrak{D}$ by explicitly specifying the argument of $f(z)$, $\arg f(z)$, at one point in $\mathfrak{D}$, and requiring that $\arg f(z)$ be defined elsewhere in $\mathfrak{D}$ by continuity. When $|\arg f(z)|<\pi$, this defines the principal branch of $\{f(z)\}^{g(z)}$.

If a_k, b_k, c_k are arbitrary complex parameters, s,t are complex variables, and p,q,m,n are integers such that $0\leq m\leq q$, $0\leq n\leq p$, we formally set

$$\Gamma_n(s+c_P+t)=\prod_{k=n+1}^{p}\Gamma(s+c_k+t),\quad \Gamma(s+c_Q+t)=\Gamma_0(s+c_Q+t),$$

$$\Gamma\left(s+c_Q^{*j}+t\right)=\prod_{\substack{k=1\\k\neq j}}^{q}\Gamma(s+c_k+t),\quad (s)_t=\frac{\Gamma(s+t)}{\Gamma(s)},\qquad (s+c_Q)_t=\prod_{k=1}^{q}(s+c_k)_t,$$

$${}_pF_q(z)={}_pF_q\left(\begin{matrix}a_P\\ b_Q\end{matrix}\middle|z\right)={}_pF_q\left(\begin{matrix}a_1,\cdots,a_p\\ b_1,\cdots,b_q\end{matrix}\middle|z\right)=\sum_{k=0}^{\infty}\frac{(a_P)_k}{(b_Q)_k}\frac{z^k}{k!},$$

$$G_{p,q}^{m,n}(w)=G_{p,q}^{m,n}\left(w\middle|\begin{matrix}a_P\\ b_Q\end{matrix}\right)=G_{p,q}^{m,n}\left(w\middle|\begin{matrix}a_1,\cdots,a_p\\ b_1,\cdots,b_q\end{matrix}\right)=\frac{1}{2\pi i}\int_L\frac{\Gamma(b_M-t)\Gamma(1-a_N+t)w^t}{\Gamma_m(1-b_Q+t)\Gamma_n(a_P-t)}dt,$$

where L is an upward oriented contour which separates the poles of $\Gamma(b_M-t)$ from those of $\Gamma(1-a_N+t)$, and which runs from $-i\infty$ to $+i\infty$ ($L=L_0$), or begins and ends at $+\infty$ ($L=L_+$), or $-\infty$ ($L=L_-$). Under lenient conditions on the parameters a_k, b_k these formal definitions yield well-defined hypergeometric functions, and Meijer *G*-functions. For example, if none of the parameters $-1+b_k$, $k=1,\cdots,q$, is a negative integer, then the formally defined ${}_pF_q(z)$ is a convergent power series when $p\leq q$, or when $p=q+1$ and $|z|<1$. If in addition to the b_k restrictions, one of the a_k's is equal to a nonpositive integer $(-m)$, then ${}_pF_q(z)$ is a polynomial in z of degree m, at most.

The basic functional relationships for the *G*-function are

$$G_{p,q}^{m,n}\left(w\middle|\begin{matrix}a_P\\ b_Q\end{matrix}\right)=G_{q,p}^{n,m}\left(w^{-1}\middle|\begin{matrix}1-b_Q\\ 1-a_P\end{matrix}\right),\qquad w^cG_{p,q}^{m,n}\left(w\middle|\begin{matrix}a_P\\ b_Q\end{matrix}\right)=G_{p,q}^{m,n}\left(w\middle|\begin{matrix}c+a_P\\ c+b_Q\end{matrix}\right).$$

*Received by the editors July 28, 1978, and in final revised form April 30, 1982. This research was sponsored by the National Research Council of Canada under grant NRC A-7549.

†Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

When the poles of the above integrand, interior to L, are simple, it follows from the residue calculus that the G-function is a finite sum of hypergeometric functions, e.g.,

$$G_{p,q}^{m,n}\left(w\,\middle|\,\begin{matrix}a_P\\ b_Q\end{matrix}\right)=\sum_{k=1}^{m}\frac{\Gamma(b_M^{*k}-b_k)\Gamma(1-a_N+b_k)}{\Gamma_m(1-b_Q+b_k)\Gamma_n(a_P-b_k)}w^{b_k}$$
$$\times{}_{p+1}F_q\left(\begin{matrix}1,1-a_P+b_k\\ 1-b_Q+b_k\end{matrix}\,\middle|\,(-1)^{p-m-n}w\right),\qquad p<q,\quad \text{or } p=q \text{ and } |w|<1.$$

Thus, $G_{p,q}^{m,n}(w)$ is a multiple-valued function, properly defined only on a Riemann surface with logarithmic branch points at $w=0,\ \infty$ and, when $p=q$, at $w=1$ or -1. This means that in any equation involving G-functions, some care must be taken to match up the proper branches of the various functions involved. For example, when $p<q$, this last equation is valid for $0<w$, $\arg w=0$, and is valid for other values of $\arg w$ by analytically continuing it with respect to w. For a more detailed discussion of hypergeometric functions and Meijer G-functions, see [7], [8].

In [6], uniform asymptotic expansions for the Meijer G-functions,

$$g_n(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G_{p+2,q}^{q,1}\left(w\,\middle|\,\begin{matrix}1-n-2\lambda,a_P,n+1\\ b_Q\end{matrix}\right),$$

$$l_{n,j}(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G_{p+3,q+1}^{q+1,2}\left(w\,\middle|\,\begin{matrix}1-n-2\lambda,a_j,a_P,n+1\\ b_Q,a_j\end{matrix}\right),\qquad j=1,\cdots,p,$$

were derived for $q-p\geq 3$, when n was large, essentially positive, and w was suitably restricted. In this paper, where $q-p=2$, Theorems 2 and 3 give similar results when w is suitably bounded away from 0 and -1 for $g_n(w)$, and from 0 for $l_{n,j}(w)$. The asymptotic expansions for $g_n(w)$ and $l_{n,j}(w)$ employ, respectively, the asymptotic scales

$$\left\{\left(\frac{|w|+1}{\sqrt{|w(w+1)(n+\lambda)^2|}}\right)^k\right\},\qquad \left\{\left(\frac{1}{w(n+\lambda)^2}\right)^k\right\}\quad \text{as } n+\lambda\to\infty,$$

which permits $|w|$ to be unboundedly large, provided $\arg w$ and $\arg(w+1)$ are suitably restricted. Theorem 8 gives pointwise asymptotic results for $g_n(e^{\pm i\pi})$. The case $q=p+2=2$ was considered by Watson [11].

When $q=p+2$, and w is suitably restricted, more explicit representations of $g_n(w)$ and $l_{n,j}(w)$ can be given. For $|\arg w|<\pi$, $|\arg(w+1)|<\pi$, we define the principal branch of $g_n(w)$ to be

$$(0.1)\qquad g_n(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\frac{1}{2\pi i}\int_{L_0}\frac{\Gamma(b_{P+2}-t)\Gamma(n+2\lambda+t)}{\Gamma(a_P-t)\Gamma(n+1-t)}w^t\,dt,$$
$$=\frac{\Gamma(n+1)\Gamma(n+2\lambda+b_{P+2})}{\Gamma(n+2\lambda)\Gamma(2n+2\lambda+1)\Gamma(n+2\lambda+a_P)}$$
$$\times w^{-n-2\lambda}{}_{p+2}F_{p+1}\left(\begin{matrix}n+2\lambda+b_{P+2}\\ 2n+2\lambda+1,n+2\lambda+a_P\end{matrix}\,\middle|\,\frac{-1}{w}\right),\qquad 1<|w|.$$

For $|\arg w|<2\pi$, we define the principal branch of $l_{n,j}(w)$ to be

(0.2)

$$l_{n,j}(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\frac{1}{2\pi i}\int_{L_0}\frac{\Gamma(b_{P+2}-t)\Gamma(1-a_j+t)\Gamma(n+2\lambda+t)}{\Gamma(a_P^{*j}-t)\Gamma(n+1-t)}w^t\,dt$$

$$=\frac{\Gamma(n+2\lambda+a_j)\Gamma(1-n-2\lambda-a_j)\Gamma(n+1)\Gamma(n+2\lambda+b_{P+2})}{\Gamma(n+2\lambda)\Gamma(2n+2\lambda+1)\Gamma(n+2\lambda+a_P)}$$

$$\times w^{-n-2\lambda}{}_{p+2}F_{p+1}\left(\begin{matrix}n+2\lambda+b_{P+2}\\ 2n+2\lambda+1, n+2\lambda+a_P\end{matrix}\middle|\frac{1}{w}\right)$$

$$+\frac{\Gamma(n+1)\Gamma(n+2\lambda-1+a_j)\Gamma(1-a_j+b_{P+2})}{\Gamma(n+2\lambda)\Gamma(n+2-a_j)\Gamma(1-a_j+a_P)}$$

$$\times w^{-1+a_j}{}_{p+3}F_{p+2}\left(\begin{matrix}1, 1-a_j+b_{P+2}\\ n+2-a_j, -n-2\lambda+2-a_j, 1-a_j+a_P\end{matrix}\middle|\frac{1}{w}\right),$$

$$1<|w|,\quad |\arg w|<\pi,$$

$$a_j-a_q\neq 0, \pm 1, \pm 2\cdots (j\neq q),\quad b_k-a_j\neq -1, -2, \cdots,$$

$$j, q=1, \cdots, p,\quad k=1, \cdots, p+2.$$

Note that the first term in the series for $l_{n,j}(w)$ is the analytic continuation of the series for $g_n(w)$ in (0.1), i.e.,

$$\Gamma(n+2\lambda+a_j)\Gamma(1-n-2\lambda-a_j)e^{\pm i\pi(n+2\lambda)}g_n(we^{\pm i\pi}).$$

Under lenient restrictions on the various parameters, the generalized Jacobi functions

$${}_{p+2}F_{p+1}\left(\begin{matrix}-n, n+2\lambda, 1-a_P\\ 1-b_{P+1}\end{matrix}\middle|w\right),\qquad n \text{ arbitrary},$$

can be written as a linear combination of the $g_n(w)$ and $l_{n,j}(w)$ when $b_{p+2}=0$. Thus, the existence of estimates for the $g_n(w)$ and $l_{n,j}(w)$ when n is large, implies the existence of similar results for the generalized Jacobi functions—see Theorems 7 and 9. In particular, when $p=0$, $b_2=0$, $b_1=-\beta$, $2\lambda=\alpha+\beta+1$, $g_n(w)$ is related to the Jacobi polynomial $P_n^{(\alpha,\beta)}(2w-1)$ in the same way that the modified Bessel function $K_\nu(z)$ is related to the Bessel function $J_\nu(z)$, i.e., in terms of the Hankel functions $H_\nu^{(1)}(z)$, $H_\nu^{(2)}(z)$,

$$J_\nu(z)=\left(\frac{1}{2}\right)\left[H_\nu^{(1)}(z)+H_\nu^{(2)}(z)\right],$$

$$H_\nu^{(j+3/2)}(z)=\left(\frac{2}{\pi}\right)e^{ij\pi(1+\nu)}K_\nu(ze^{ij\pi}),\qquad j=\pm\frac{1}{2},$$

whereas

$$P_n^{(\alpha,\beta)}(2w-1)=\frac{(-1)^n(1+\beta)_n}{n!}{}_2F_1\left(\begin{matrix}-n, n+\alpha+\beta+1\\ 1+\beta\end{matrix}\middle|w\right)$$

$$=\frac{(-1)^n\Gamma(n+1+\beta)}{2\pi n!}\left[e^{i\pi(1/2+\beta)}g_n(we^{i\pi})+e^{-i\pi(1/2+\beta)}g_n(we^{-i\pi})\right].$$

For the Jacobi polynomials $P_n^{(\alpha,\beta)}(2w-1)$, our asymptotic expansion agrees with the classical result of Darboux, see Szego [10,(8.21.18)], but is valid in a wider sector than previously recorded for the classical expansion.

The results of Theorems 2 and 3 can also be used to derive rational approximations to certain Meijer G-functions, see [5].

1. An integral representation for $g_n(w)$. The asymptotic expansion of $g_n(w)$ is derived in Theorem 2 from the following integral representation, and its extension in Corollary 1.1.

THEOREM 1. *Let $p+1$ be a positive integer; n,λ,a_j,b_j be complex constants such that $n\neq$ a negative integer; $\sigma=\frac{1}{2}-2\lambda+\sum_{j=1}^{p}a_j-\sum_{j=1}^{p+2}b_j\neq$ an integer;*

$$\operatorname{Re}(n+2\lambda)>0, \qquad \operatorname{Re}(n+2\lambda+b_j)>0, \qquad j=1,\cdots,p+2.$$

Next, let w be a complex variable restricted to the domain $\mathfrak{D}_w^0=(0,\infty)$, where $\arg w=\arg(1+w)=0$, and

$$\xi=\xi(w)=\log\left\{\sqrt{w}+\sqrt{1+w}\right\}, \qquad \arg\xi=0 \quad \text{if } w\in\mathfrak{D}_w^0, \tag{1.1}$$

so that $w\in\mathfrak{D}_w^0$, implies $\xi>0$, and $e^{\xi}>1$.

Finally, let $\mathbb{S}(z)$ be defined initially by

$$\mathbb{S}(z)=\Gamma(\sigma+1)G_{p+2,p+2}^{p+2,0}\left(z\,\middle|\,\begin{matrix}a_P,\frac{1}{2}-\lambda,1-\lambda\\ b_{P+2}\end{matrix}\right),$$

$$0<|z|<1, \quad |\arg z|<\pi, \quad |\arg(1-z)|<\pi.$$

As $\mathbb{S}(z)$ is a solution of $\mathfrak{L}y=0$,

$$\mathfrak{L}=\prod_{j=1}^{p+2}(\delta-b_j)-z(\delta+\lambda)\left(\delta+\lambda+\frac{1}{2}\right)\prod_{j=1}^{p}(\delta+1-a_j), \qquad \delta=z\frac{d}{dz},$$

a $(p+2)$nd order linear differential equation whose only singularities are the regular singular points $z=0,1$ and ∞, $\mathbb{S}(z)$ can be analytically continued outside $|z|<1$, along any arc avoiding these singular points. In particular, in the fundamental neighborhood near $z=1$,

$$\mathfrak{D}=\left\{z:\begin{matrix}(-\infty,1] \text{ a branch cut}, 0<|z-1|<1,\\ |\arg z|<\pi, |\arg(z-1)|<\pi, \arg(1-z)=\pi+\arg(z-1)\end{matrix}\right\},$$

Nørlund [9] *has shown that $\mathbb{S}(z)$ has the local representation*

$$\mathbb{S}(z)=\mathbb{S}_p(z)=(1-z)^{\sigma}\mathbb{S}_p^*(z)=\left[e^{i\pi}(z-1)\right]^{\sigma}\mathbb{S}_p^*(z), \qquad z\in\mathfrak{D}, \tag{1.2}$$

$$\mathbb{S}_p^*(z)=\sum_{j=0}^{\infty}d_j\frac{(1-z)^j}{(1+\sigma)_j}, \qquad d_0=1, \quad |1-z|<1, \quad |\arg z|<\pi, \tag{1.3}$$

which we will denote as the principal branch of $\mathbb{S}(z)$, where the constants d_j can be computed recursively from $\mathfrak{L}\mathbb{S}_p(z)=0$. By continuity, $\mathbb{S}_p(z)$ and $\mathbb{S}_p^*(z)$ are defined on $\overline{\mathfrak{D}}$, the closure of $\mathfrak{D}$. In fact, $\mathbb{S}_p^*(z)$ can be analytically continued outside of $\overline{\mathfrak{D}}$ to be a globally analytic function function $\mathbb{S}^*(z)$ on a Riemann surface whose only singularities are logarithmic branch points at $z=0,1$ and ∞. Thus, when $z\in\overline{\mathfrak{D}}$, $\mathbb{S}^*(z)$ has the local representation $\mathbb{S}_p^*(z)$, which we will denote as the principal branch of $\mathbb{S}^*(z)$.

Then for $w\in\mathcal{D}_w^0$,

$$(1.4)\quad g_n(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G^{p+2,1}_{p+2,p+2}\left(w\,\middle|\,\begin{matrix}1-n-2\lambda,a_p,n+1\\ b_{p+2}\end{matrix}\right)$$

$$=\sqrt{\pi}\,4^{\lambda}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\frac{\Gamma(-\sigma)}{2\pi i}\int_{+\infty}^{e^{2\xi}-}\frac{(t-1)^{-2\lambda}}{t^{n+1}}e^{-i\pi\sigma}\mathcal{S}\left(\frac{4wt}{(t-1)^2}\right)dt,$$

where $\mathcal{C}^*$, *the loop contour of integration and the phases of the various factors in the integrand of* (1.4) *are indicated in Fig.* 1. *Note that the t-plane has a branch cut along* $[e^{2\xi},\infty)$, *and that* $\mathcal{C}^*$ *starts/stops at* $+\infty$, *encloses* $t=e^{2\xi}$ *locally with the negatively oriented circle* $|t-e^{2\xi}|=\rho^*$, $0<\rho^*<|e^{2\xi}-1|$, *but encloses none of the other singularities* $t=0,1$ *and* $e^{-2\xi}$.

Also, for $w\in\mathcal{D}_w^0$, $t\in\mathcal{C}^*$, ρ^* *sufficiently small, and*

$$(1.5)\qquad X=\frac{4wt}{(t-1)^2}\quad or\quad 1-X=\frac{(t-e^{2\xi})(t-e^{-2\xi})}{(t-1)^2},$$

a simple computation shows that $X\in\overline{\mathcal{D}}$, *and hence that* $e^{-i\pi\sigma}\mathcal{S}(X)$ *in* (1.4) *reduces to*

$$e^{-i\pi\sigma}\mathcal{S}(X)=\left\{\frac{e^{-i\pi}(t-e^{2\xi})(t-e^{-2\xi})}{(t-1)^2}\right\}^{\sigma}\mathcal{S}_p^*(X).$$

In particular, the argument of $X-1=e^{-i\pi}(1-X)$ *is zero at the point* Q^* *in Fig.* 1.

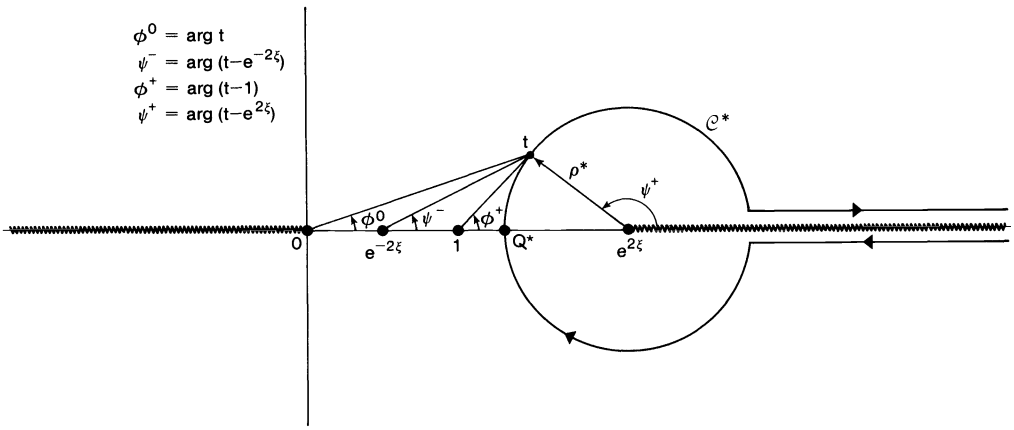


FIG 1. *t-plane,* wwww *a branch cut.*

Proof. As $g_n(w)$ and $\mathcal{S}(X)$ are multiple-valued functions, they are properly defined only on Riemann surfaces with logarithmic branch points at $w=0,-1,\infty$ and $t=0,1,e^{\pm 2\xi},\infty$, respectively, and extreme care must be taken to match up the proper branches of $g_n(w)$ and $\mathcal{S}(X)$ in the integral representation (1.4). For $w\in\mathcal{D}_w^0$, we take for $g_n(w)$ the principal value defined in (0.1), while for $\mathcal{S}(X)$, we will take the principal branch $\mathcal{S}_p(X)$. A proof of Theorem 1 can be constructed which exploits the analytic continuation properties of $\mathcal{S}(X)$, as developed by Nørlund in [9], to explicitly evaluate the integral in the theorem as the series in (0.1). Alternately, we give the following proof of Theorem 1 which was communicated to us by the referee.

For $w\in\mathfrak{D}_w^0$, tentatively assume

$$\text{(1.6)}\qquad \operatorname{Re}\left(n+\lambda+\frac{1}{2}\right)>0, \qquad \operatorname{Re}\sigma>0,$$

and choose κ such that

$$\kappa<\frac{1}{2}-\operatorname{Re}\lambda, \qquad -\operatorname{Re}(n+2\lambda)<\kappa<\operatorname{Re}b_j, \qquad j=1,\cdots,p+2.$$

Then by definition,

(1.7)

$$g_n(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\cdot\frac{1}{2\pi i}\int_{\kappa-i\infty}^{\kappa+i\infty}\frac{\Gamma(n+2\lambda+s)\Gamma(\frac{1}{2}-\lambda-s)\Gamma(1-\lambda-s)}{\Gamma(n+1-s)}f(s)w^s\,ds,$$

where

$$f(s)=\frac{\Gamma(b_{P+2}-s)}{\Gamma(\frac{1}{2}-\lambda-s)\Gamma(1-\lambda-s)\Gamma(a_P-s)},$$

$$=(-s)^{-\sigma-1}\{1+\mathcal{O}(s^{-1})\}, \qquad s\to\infty, \quad \operatorname{Re}s\leq \text{ a fixed number } K,$$

and the integration contour separates the poles of $\Gamma(n+2\lambda+s)$ from those of $\Gamma(\frac{1}{2}-\lambda-s)\Gamma(1-\lambda-s)\Gamma(b_{P+2}-s)$. From the beta integral, we have for $\operatorname{Re}s=\kappa$,

$$\frac{\Gamma(n+2\lambda+s)\Gamma(\frac{1}{2}-\lambda-s)\Gamma(1-\lambda-s)}{\Gamma(n+1-s)}=\sqrt{\pi}\,4^{\lambda+s}\int_1^\infty t^{s-n-1}(t-1)^{-2\lambda-2s}dt.$$

As $\operatorname{Re}\sigma>0$, and $\kappa+\operatorname{Re}(n+2\lambda)>0$, the last integral can be substituted into (1.7) and the order of integration interchanged. This process yields the result

$$\text{(1.8)}\qquad g_n(w)=\sqrt{\pi}\,4^\lambda\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_1^\infty\frac{(t-1)^{-2\lambda}}{t^{n+1}}\frac{1}{2\pi i}\int_{\kappa-i\infty}^{\kappa+i\infty}X^s f(s)\,ds\,dt,$$

where X has the same meaning as in (1.5).

Applying the residue calculus to the Mellin–Barnes integral in (1.8), we see that when $\arg X=\arg(1-X)=0$,

$$\frac{1}{2\pi i}\int_{\kappa-i\infty}^{\kappa+i\infty}X^s f(s)\,ds=\begin{cases}G_{p+2,p+2}^{p+2,0}\left(X\,\middle|\,\begin{matrix}a_P,\frac{1}{2}-\lambda,1-\lambda\\ b_{P+2}\end{matrix}\right)=\dfrac{\mathfrak{S}_p(X)}{\Gamma(\sigma+1)} & \text{if } 0<X<1,\\ 0 \quad \text{if } 1<X.\end{cases}$$

Moreover, for $\xi>0$ and $t=\exp\{2\xi+2v\}$, it follows that

$$\text{(1.9)}\qquad 0<X=\frac{4wt}{(t-1)^2}=\frac{\sinh^2\xi}{\sinh^2(\xi+v)}\begin{cases}<1 & \text{if } e^{2\xi}<t,\\ >1 & \text{if } 1<t<e^{2\xi}.\end{cases}$$

Thus, in (1.8), the t interval of integration can be replaced by $[e^{2\xi},\infty)$, and the last factor of the integrand can be identified with $\mathfrak{S}_p(X)/\Gamma(\sigma+1)$, i.e.,

$$\text{(1.10)}\qquad g_n(w)=\frac{\sqrt{\pi}\,4^\lambda\Gamma(n+1)}{\Gamma(2n+\lambda)\Gamma(\sigma+1)}\int_{e^{2\xi}}^\infty\frac{(t-1)^{-2\lambda}}{t^{n+1}}\mathfrak{S}_p\left(\frac{4wt}{(t-1)^2}\right)dt.$$

Theorem 1 is just a contour integral formulation of (1.10) which makes use of Nørlund's results (1.2). To see this explicitly, let I denote the last half-line of (1.4). As $\mathrm{Re}\,\sigma>0$, we can take $\rho^*=0$ in Fig. 1, to shrink $\mathcal{C}^*$ in I into the two straight lines $\mathcal{L}^+$ and $\mathcal{L}^-$ along the upper and lower edges of the branch cut $[e^{2\xi},\infty)$, respectively. From (1.9), it follows that $t\in\mathcal{L}^+\cup\mathcal{L}^-$ implies $0<X\leq 1$. Also, using the integration variable $v\in[e^{2\xi},\infty)$, $\arg v=\arg(1+v)=0$, we have that $t\in\mathcal{L}^+$ implies $\arg X=\arg(1-X)=0$, $X\in\mathcal{D}$ and

$$e^{-i\pi\sigma}\mathcal{S}\left(\frac{4wt}{(t-1)^2}\right)=e^{-i\pi\sigma}\mathcal{S}_p\left(\frac{4wv}{(v-1)^2}\right),$$

while $t\in\mathcal{L}^-$ implies $\arg X=0$, $\arg(1-X)=2\pi$, $X\in\overline{\mathcal{D}}$ and

$$e^{-i\pi\sigma}\mathcal{S}\left(\frac{4wt}{(t-1)^2}\right)=e^{i\pi\sigma}\mathcal{S}_p\left(\frac{4wv}{(v-1)^2}\right).$$

Combining these results, we can write

$$I=\sqrt{\pi}\,4^{\lambda}\frac{\Gamma(n+1)\Gamma(-\sigma)}{\Gamma(n+2\lambda)}\,\frac{[e^{-i\pi\sigma}-e^{i\pi\sigma}]}{2\pi i}\int_{e^{2\xi}}^{\infty}\frac{(v-1)^{-2\lambda}}{v^{n+1}}\mathcal{S}_p\left(\frac{4wv}{(v-1)^2}\right)dv=g_n(w),$$

in view of (1.10). By analytic continuation with respect to the various parameters, (1.4) remains true when the tentative assumptions (1.6) are relaxed.

As the other results in Theorem 1 follow from straightforward geometrical arguments which are left to the reader, this completes Theorem 1.

Remark 1. The d_j in (1.3) satisfy a linear difference equation of length $(p+3)$. The first few d_j are as follows. Letting

$$(x+1-\lambda)\left(x+\frac{1}{2}-\lambda\right)\prod_{j=1}^{p}(x+a_j)=\sum_{j=0}^{p+2}A_jx^{p+2-j},$$

$$\prod_{j=1}^{p+2}(x+B_j)=\sum_{j=0}^{p+2}B_jx^{p+2-j},\qquad C_j=B_j-A_j,$$

we have by direct computation

$$\begin{aligned}
d_1&=B_1C_1-C_2,\\
2d_2&=[(1-B_1)(1-C_1)-C_2]d_1+(B_2C_1-C_3),\\
3d_3&=[(2-B_1)(2-C_1)-C_2]d_2+[B_2(C_1-1)-C_3]d_1\\
&\qquad+(1-C_1)(B_2C_1-C_3)+(B_3C_1-C_4).
\end{aligned}$$

Note that $\sigma=-1-C_1$.

Remark 2. Consider the variable ξ defined in (1.1). Simple computations show that for $w\in\mathcal{D}_w^0$,

$$e^{\xi}=\sqrt{1+w}+\sqrt{w},\quad \sinh\xi=\sqrt{w},\quad \sinh 2\xi=2\sqrt{w(1+w)},$$

$$e^{-\xi}=\sqrt{1+w}-\sqrt{w},\quad \cosh\xi=\sqrt{1+w},\quad \cosh 2\xi=1+2w,$$

and that if we make the change of integration variable $t=\exp\{2\xi+2ue^{i\pi}\}$ in (1.4), we obtain

$$
(1.11)\quad \begin{cases} X=\dfrac{4wt}{(t-1)^2}=\left(\dfrac{\sinh(\xi-u)}{\sinh\xi}\right)^{-2}, \\ X-1=\dfrac{e^{-i\pi}(t-e^{2\xi})(t-e^{-2\xi})}{(t-1)^2}=\dfrac{(\sinh u)\sinh(2\xi-u)}{[\sinh(\xi-u)]^2}, \\ g_n(w)=\sqrt{\pi}\,(\sinh\xi)^{-2\lambda-\sigma}(\cosh\xi)^{\sigma}\dfrac{\Gamma(n+1)}{\Gamma(n+2\lambda)} \\ \qquad\times\dfrac{\Gamma(-\sigma)}{2\pi i}\displaystyle\int_{\mathcal{C}_0} e^{2(u-\xi)(n+\lambda)}(2u)^{\sigma}X^{\lambda}\left\{\frac{X-1}{2u\coth\xi}\right\}^{\sigma}\mathbb{S}_p^*(X)\,d(2u), \end{cases}
$$

where $\mathcal{C}_0$ is an infinite loop contour in the u-plane, starting/stopping at $-\infty$, enclosing $u=0$ with the positively oriented circle $|u|=\rho$, but enclosing none of the other singularities of the integrand, i.e., the points $u=k\xi+i\pi q$, $k=0,1$ or 2, and q an integer. The u-plane is chosen to have $|\arg u|\le\pi$, and to have branch cuts as indicated in Fig. 2. This copy of the u-plane will be denoted by $\mathfrak{U}$, as long as $\operatorname{Re}\xi\ge 0$.

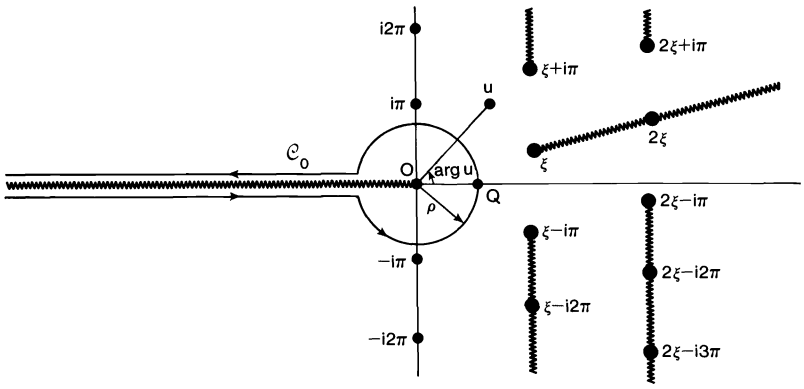


FIG 2. $\mathfrak{U}$-*surface*, $\operatorname{Re}\xi\ge 0$, $Q\in\mathfrak{D}_\xi^0$, ⟿ *a branch cut.*

If the complex variable ξ is restricted to

$$\mathfrak{D}_\xi^0=\{\xi>0:\ |\arg(\xi+iq\pi/2)|\le\pi/2,\ q \text{ an integer}\},$$

then there is a 1-1 correspondence between $\mathfrak{D}_w^0$ and $\mathfrak{D}_\xi^0$. In fact, if $\mathfrak{D}_\xi^0$ is contained in $\mathfrak{D}_\xi^{\#}$, a simply connected region located on a Riemann surface with logarithmic branch points at $\xi=\infty$ and $\xi=iq\pi/2$, q an integer, but with $\mathfrak{D}_\xi^{\#}$ not containing any of these singular points, and $\mathfrak{D}_w^{\#}$ is the image of $\mathfrak{D}_\xi^{\#}$ under the mapping $w=\sinh^2\xi$, then the functions

$$\xi=\xi(w)=\log\left[\sqrt{w}+\sqrt{1+w}\right],\qquad \arg\xi=0 \quad \text{if } w\in\mathfrak{D}_w^0,$$

$$w=w(\xi)=\sinh^2\xi,\qquad \arg w=\arg(1+w)=0 \quad \text{if } \xi\in\mathfrak{D}_\xi^0$$

are inverse, conformal mappings to each other on $\mathfrak{D}_\xi^{\#}$ and $\mathfrak{D}_w^{\#}$. Some of the properties of such $\mathfrak{D}_\xi^{\#}$, $\mathfrak{D}_w^{\#}$ sets are indicated in Figs. 3 and 4. Clearly, ξ is purely imaginary, if and only if $\arg w - \arg(1+w) = \pi \pmod{2\pi}$, which, in turn, is true, if and only if $w \in [-1, 0]$. The singular points $\xi = iq\pi/2$, q an integer, correspond to the points $w = 0$, or -1, as q is even or odd, respectively, and as the $iq\pi/2 \notin \mathfrak{D}_\xi^{\#}$, the points $0, -1 \notin \mathfrak{D}_w^{\#}$. For more detail, fix $\xi_0 \in \mathfrak{D}_\xi^0$. Then as ξ varies along a straight line connecting ξ_0 to $\xi_q = \xi_0 + iq\pi/2$ on the ξ-surface, w starts on the w-surface at $w_0 = \sinh^2 \xi_0$ and makes q half-revolutions around $w = 0$ and -1 without intersecting $[-1, 0]$. If $w_q = \sinh^2 \xi_q$, it follows from continuity considerations that

$$\arg w_q = \arg(1 + w_q) = q\pi, \qquad |2^{-1}[1 + (-1)^{q+1}] + w_q| = w_0.$$

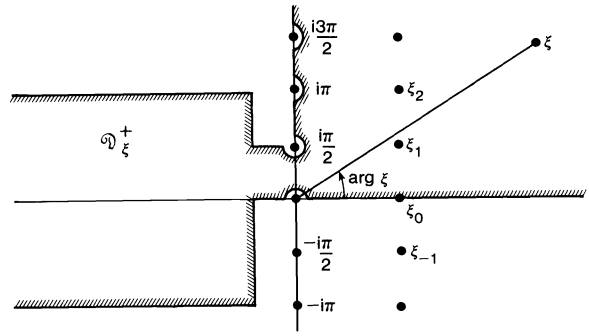


FIG 3. *ξ-surface.*

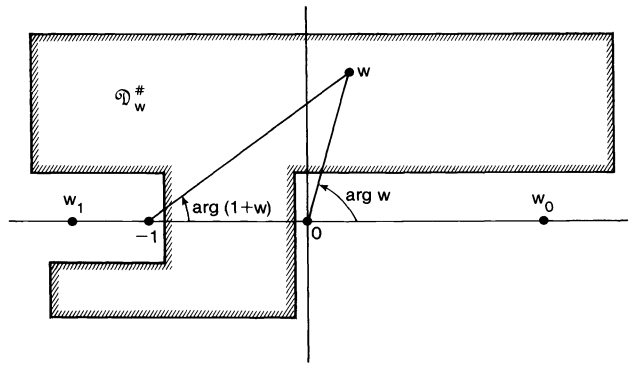


FIG 4. *w-surface.*

We now generalize Theorem 1 to complex values of w and ξ.

COROLLARY 1.1. *Unless modified explicitly, the notation and restrictions of Theorem 1 will be assumed. On a Riemann surface $\mathfrak{R}_\xi$ with branch points at $\xi = \infty$ and $\xi = iq\pi/2$, q an integer, let $\mathfrak{D}_\xi^{\#}$ be the simply connected region defined as follows.*

$$\mathfrak{D}_\xi^{\#} = \mathfrak{D}_\xi^+ \cup \mathfrak{D}_\xi^-,$$

$$\mathfrak{D}_\xi^0 = \mathfrak{D}_\xi^+ \cap \mathfrak{D}_\xi^- = \{\xi > 0 : |\arg(\xi + iq\pi/2)| \leq \pi/2, \, q \text{ an integer}\},$$

where $\mathcal{D}_\xi^-$ *is the complex conjugate of* $\mathcal{D}_\xi^+$, *and* $\mathcal{D}_\xi^+$ *is the unbounded, simply connected set defined by*

$$(1.12)\quad \mathcal{D}_\xi^+ = \left\{ \xi=\alpha+i\beta: \begin{array}{ll} \alpha>0 & \Rightarrow \beta\geq 0, \\ \alpha=0+ & \Rightarrow \beta>0,\ \beta\neq q\pi/2,\ q \text{ an integer}, \\ -\delta\leq\alpha\leq 0- & \Rightarrow 0<\beta<\pi/2, \\ \alpha\leq -\delta & \Rightarrow |\beta|\leq\kappa,\ \delta,\kappa \text{ fixed positive numbers} \end{array} \right\},$$

and represented in Fig. 3. *Also, let* $\mathcal{D}_w^\#$, *the image of* $\mathcal{D}_\xi^\#$ *under the mapping* $w=\sinh^2\xi$, *be the corresponding simply connected region on a Riemann surface* $\mathcal{R}_w$ *with logarithmic branch points at* $w=0,-1$ *and* ∞.

Then for $\xi=\alpha+i\beta\in\mathcal{D}_\xi^\#$, *or* $w\in\mathcal{D}_w^\#$,

$$(1.13)\quad g_n(w)=\sqrt{\pi}\,(\sinh\xi)^{-2\lambda-\sigma}(\cosh\xi)^\sigma\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)} \times\frac{\Gamma(-\sigma)}{2\pi i}\int_{\mathcal{C}} e^{2(u-\xi)(n+\lambda)}(2u)^\sigma K(u,\xi)\,d(2u),$$

$$(1.14)\quad K(u,\xi)=X^\lambda\left\{\frac{X-1}{2u\coth\xi}\right\}^\sigma \mathbb{S}^*(X),\qquad u\in\mathcal{C},\quad \xi\in\mathcal{D}_\xi^\#,$$

$$X=\left\{\frac{\sinh(\xi-u)}{\sinh\xi}\right\}^{-2},\qquad \frac{X-1}{2u\coth\xi}=\frac{(\sinh u)\sinh(2\xi-u)}{2u(\coth\xi)[\sinh(\xi-u)]^2},$$

where $\mathcal{C}$ *lies on the sheet* $\mathfrak{A}$ *of* $\mathcal{R}_\xi$, *and is an infinite loop contour starting/ending at* $-\infty$, *enclosing* $u=0$ *within the positively oriented circle* $|u|=\rho$, *but enclosing none of the other singularities of* $(2u)^\sigma K(u,\xi)$, *i.e., the points* $u=k\xi+i\pi q$, $k=0,1$ *or* 2, *and* q *an integer. The branch cuts in* $\mathfrak{A}$ *depend upon the choice of* $\mathcal{C}$. *For* $\alpha\geq 0$, *we choose* $\mathcal{C}=\mathcal{C}_0$, *and the branch cuts as indicated in Fig.* 2.

For $\alpha<0$, *we choose* $\mathcal{C}=\mathcal{C}_1+\mathcal{L}_1+\mathcal{L}_2$, *a contour of the general form shown in Fig.* 5, *together with its attendant branch cuts. In some special cases for* $\alpha<0$, $\mathcal{C}=\mathcal{C}_0$, *see Fig.* 6. *As point sets,*

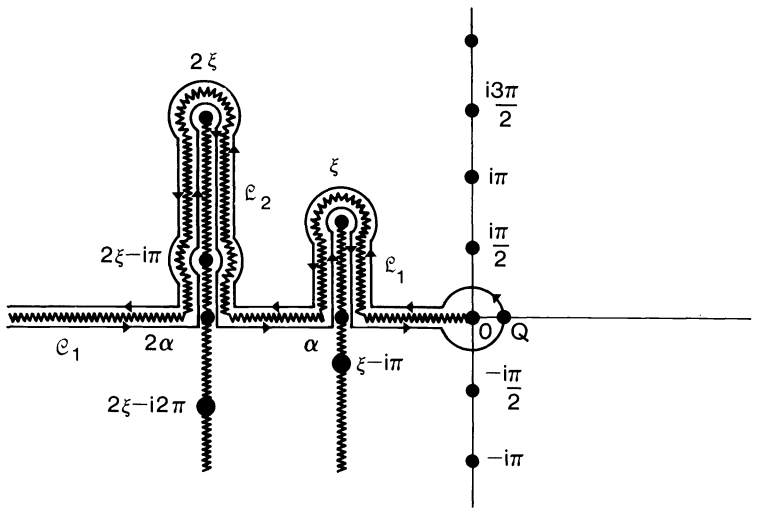


FIG 5. $\mathfrak{A}$*-surface*, $\mathrm{Re}\,\xi<0$, $\xi\in\mathcal{D}_\xi^-$, $Q\in\mathcal{D}_\xi^0$, wwwww *branch cuts.*

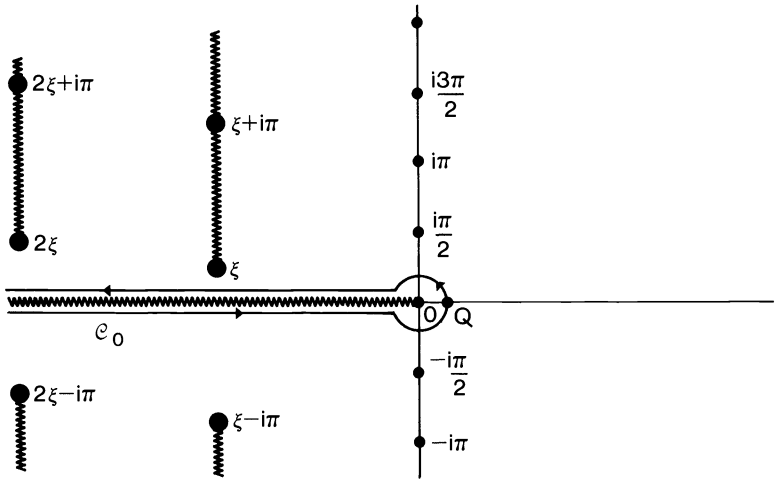


FIG 6. $\mathfrak{A}$-*surface*, $\operatorname{Re}\xi<0$, $\xi\in\mathfrak{D}_\xi^+$, $0<\operatorname{Im}\xi<\frac{\pi}{2}$, $Q\in\mathfrak{D}_\xi^0$, wwww *branch cuts.*

$$\begin{aligned}(1.15)\qquad &\mathcal{C}_0=\{u=e^{\pm i\pi}x;\rho\le x\}\cup\{u=\rho e^{i\psi}:-\pi\le\psi\le\pi\},\\ &\mathcal{C}_1\subset\mathcal{C}_0,\ \mathcal{L}_j=\mathcal{L}_j^*\cup\mathcal{L}_j^{**},\ j=1\text{ or }2,\\ &\mathcal{L}_j^*\subset\{u=j\alpha+i\tau:|\tau|\le j\kappa,\ \kappa\text{ defined in (1.12)}\},\\ &\mathcal{L}_j^{**}\subset\{u=j\xi+i\pi q+\rho e^{i\psi}:-\pi\le\psi\le\pi,\ q\text{ an integer}\}.\end{aligned}$$

In all cases, $\mathcal{C}$ *crosses the positive real axis of* $\mathfrak{A}$ *at* $Q\in\mathfrak{D}_\xi^0$, *so that* $\arg u=0$ *at* Q. *Elsewhere in* $\mathfrak{A}$, $\arg u$ *is defined by continuity.*

As $u\in\mathcal{C}_0$ *and* $\xi\in\mathfrak{D}_\xi^0$ *imply that* $X\in\mathfrak{D}$, *the multiple-valued function* $K(u,\xi)$ *is specified to have the well-defined local representation*

$$\begin{aligned}K(u,\xi)&=K_p(u,\xi)\\ &=X^\lambda\left\{\frac{X-1}{2u\coth\xi}\right\}^\sigma\mathbb{S}_p^*(X)=X^\lambda\left\{\frac{e^{-i\pi}}{2u\coth\xi}\right\}^\sigma\mathbb{S}_p(X),\qquad u\in\mathcal{C}_0,\quad\xi\in\mathfrak{D}^0,\end{aligned}$$

which we will denote as the principal branch of $K(u,\xi)$. *Elsewhere, for* $u\in\mathcal{C}$ *and* $\xi\in\mathfrak{D}^\#$, $K(u,\xi)$ *is defined to be* $K_p(u,\xi)$ *analytically continued with respect to* u *and* ξ. *This same analytic continuation procedure applied to* $g_n(w)=g_n(\sinh^2\xi)$ *determines the branch of* $g_n(w)$ *which occurs in* (1.13), *and when applied to the functions* $\mathbb{S}_p(X)$, $(1-X)^\sigma=[e^{i\pi}(X-1)]^\sigma$, $\mathbb{S}_p^*(X)$, *for* $X\in\mathfrak{D}$, *yields the analytically continued form of* (1.2), *i.e., the basic relationship*

$$\mathbb{S}(X)=(1-X)^\sigma\mathbb{S}^*(X)=\left[e^{i\pi}(X-1)\right]^\sigma\mathbb{S}^*(X),\qquad u\in\mathcal{C},\quad\xi\in\mathfrak{D}^\#;$$

in particular, this determines the branch of $\mathbb{S}^*(X)$ *which occurs in* (1.14).

Proof. For $\xi\in\mathfrak{D}^0$, $\mathcal{C}$ reduces to $\mathcal{C}_0$ and the corollary reduces to the alternate form of Theorem 1 stated in (1.11) of Remark 2. The remainder of the corollary follows by analytic continuation with respect to u and ξ.

Remark 3. In the proof of Theorem 2, it is actually shown that, for $\xi\in\mathfrak{D}_\xi^\#$, $K(u,\xi)$ can be analytically continued as a function of u into a neighborhood of $u=0$ which depends on ξ. Using the Maclaurin series expansion of $K(u,\xi)$ at $u=0$ as the definition of $K(u,\xi)$, $K(u,\xi)$ could then be defined on $\mathcal{C}$ by analytic continuation along $\mathcal{C}$.

This would define $K(u,\xi)$ independently of any explicit mention of the multiple valued functions $\mathbb{S}(X)$ and $\mathbb{S}^*(X)$.

Remark 4. In the special case $p=0$, $b_2=0$, $b_1=b$, $\sigma=\frac{1}{2}-2\lambda-b$,

$$\mathbb{S}(z)=(1-z)^{\sigma}{}_2F_1\left(\begin{matrix}\sigma+\lambda,\sigma+\lambda+\frac{1}{2}\\ \sigma+1\end{matrix}\middle|1-z\right)=z^b(1-z)^{\sigma}{}_2F_1\left(\begin{matrix}1-\lambda,\frac{1}{2}-\lambda\\ \sigma+1\end{matrix}\middle|1-z\right).$$

If in addition, $\lambda=\frac{1}{2}$, then $\sigma=-\frac{1}{2}-b$, $\mathbb{S}(z)=z^b(1-z)^{\sigma}$, and Theorem 1, Corollary 1.1 reduce to

$$\begin{aligned}g_n(w)&=\frac{\Gamma(n+1)\Gamma(n+1+b)}{\Gamma(2n+2)w^{n+1}}{}_2F_1\left(\begin{matrix}n+1,n+1+b\\ 2n+2\end{matrix}\middle|\frac{-1}{w}\right)\\&=\sqrt{\pi}\,4^{-\sigma}e^{-i\pi\sigma}\frac{\Gamma(-\sigma)}{2\pi i}\int_{+\infty}^{(e^{2\xi}-)}\frac{(wt)^b}{t^{n+1}}\{(t-e^{2\xi})(t-e^{-2\xi})\}^{\sigma}dt\\&=\sqrt{\pi}\,w^{(-1+2b)/4}(1+w)^{(-1-2b)/4}e^{-(2n+1)\xi}\\&\quad\times\frac{\Gamma(-\sigma)}{2\pi i}\int_{-\infty}^{(0+)}e^{u(2n+1)}(2u)^{\sigma}\left\{\left(\frac{\sinh u}{u}\right)\left(\frac{\sinh(2\xi-u)}{\sinh 2\xi}\right)\right\}^{\sigma}d(2u),\end{aligned}$$

which is equivalent to an integral given by Watson [11, §13] when $b=0$.

2. The asymptotic expansion of $g_n(w)$ and $I_{n,j}(w)$.

THEOREM 2. *Let $p,n,\lambda,a_j,b_j,\sigma,d_j,w,\xi=\alpha+i\beta$ (α,β real) and $\mathcal{D}_{\xi}^{\#}$ be as in Theorem 1 and Corollary* 1.1, *with the additional restrictions that n be a large parameter such that $|n|\to+\infty$, $\arg n=\mathcal{O}(n^{-1})$ as $n\to\infty$, and λ,a_j,b_j be bounded with respect to n. Set*

$$\begin{aligned}&n^*=n+\lambda,\quad \text{with the general restriction } |n^*|\geq e^2,\\(2.1)\qquad&\omega=\frac{1+|w|}{2\sqrt{|w(1+w)|}}=\frac{1+|\sinh\xi|^2}{|\sinh 2\xi|},\end{aligned}$$

and let $\mathcal{D}_{\xi}$ be a subset of $\mathcal{D}_{\xi}^{\#}$ which depends on n, and satisfies the following restrictions:

(i) $(32)^{-1}>|\omega/n^*|=o(1)$, *as $n\to\infty$, uniformly for $\xi\in\mathcal{D}_{\xi}$.*

(ii) *If $\xi=\alpha+i\beta\in\mathcal{D}_{\xi}$, $\alpha<0$, and $|\beta|\notin(0,\pi/2)$, then*

$$|\alpha n^*|>\log|n^*|=o(\alpha n^*)\quad \text{as } n\to\infty.$$

We take $\mathcal{D}_w$ to be the image of $\mathcal{D}_{\xi}$ under the mapping $w=\sinh^2\xi$. In particular, $\mathcal{D}_{\xi}$ can be chosen such that $\mathcal{D}_w$ contains the domain

$$\mathcal{D}_w^*=\left\{\begin{matrix}w:\ |\arg w|\leq 3\pi-\varepsilon_1,\ |w|\geq|n^*|^{-1}\{\log|n^*|\}^{1+\varepsilon_3},\\ |\arg(1+w)|\leq 2\pi-\varepsilon_2,\ |1+w|\geq|n^*|^{-1}\{\log|n^*|\}^{1+\varepsilon_4}\end{matrix}\right\},$$

where the ε_j are small positive numbers independent of n.

Then for all $\xi\in\mathcal{D}_{\xi}$ or $w\in\mathcal{D}_w$, there exist functions $S_k(\xi)$ such that for $m+1$ an arbitrary positive integer,

$$\begin{aligned}(2.2)\qquad g_n(w)&=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G_{p+2,p+2}^{p+2,1}\left(w\middle|\begin{matrix}1-n-2\lambda,a_P,n+1\\ b_{P+2}\end{matrix}\right),\\&=\sqrt{\pi}\left[(n+\lambda)^2w\right]^{\gamma}(1+w)^{-\gamma-\lambda}e^{-2(n+\lambda)\xi}S(n+\lambda,\xi),\end{aligned}$$

(2.3)

$$S(n+\lambda,\xi)=1+\sum_{k=1}^{m-1} S_k(\xi)(n^*)^{-k}+\left(\frac{\omega}{n^*}\right)^m \mathcal{O}(1), \qquad n+\lambda\to\infty,$$

$$2\gamma=-\frac{1}{2}+\sum_{j=1}^{p+2} b_j-\sum_{j=1}^{p} a_j, \qquad \sigma=-2\lambda-2\gamma,$$

$$S_k(\xi)=\sum_{j=0}^{[k/2]} P_{k-2j,k}(\coth\xi)^{k-2j}+\sum_{j=0}^{[(k-1)/2]} Q_{k-2j,k}(\coth 2\xi)^{k-2j},$$

$$P_{11}=d_1-(\sigma+1)(\sigma+\lambda),$$

$$4P_{22}=4d_2-(\sigma+2)(3\sigma+4\lambda+3)d_1+(\sigma+1)(\sigma+2)(\sigma+\lambda)(\sigma+2\lambda+1),$$

$$12P_{02}=(\sigma+1)(\sigma+2)\{3d_1-[3\sigma^2+(3\lambda+1)\sigma+3\lambda]\}+2\lambda(\lambda-1)(2\lambda-1),$$

$$\begin{aligned}96P_{33}=&96d_3-24(\sigma+3)(3\sigma+4\lambda+6)d_2\\&+3(\sigma+2)(\sigma+3)[9\sigma^2+(24\lambda+25)\sigma+16(\lambda+1)^2]d_1\\&-(\sigma+1)_3(\sigma+\lambda)[7\sigma^2+(20\lambda+15)\sigma+8(\lambda+1)(2\lambda+1)],\end{aligned}$$

$$\begin{aligned}96P_{13}=&(\sigma+2)(\sigma+3)\{24d_2-[21\sigma^2+(24\lambda+53)\sigma+(48\lambda+32)]d_1\\&+(\sigma+1)(\sigma+\lambda)[9\sigma^2+(12\lambda+17)\sigma+(24\lambda+8)]\}\\&+16\lambda(\lambda-1)(2\lambda-1)[d_1-(\sigma+1)(\sigma+\lambda)],\end{aligned}$$

$$2Q_{11}=\sigma(\sigma+1), \quad 8Q_{22}=(\sigma-1)_4, \quad 48Q_{33}=(\sigma-2)_6,$$

$$48Q_{13}=(\sigma)_4\{3d_1-[3\sigma^2+(3\lambda-1)\sigma+(3\lambda+2)]\}+4\lambda(\lambda-1)(2\lambda-1)(\sigma)(\sigma+1), \quad etc.$$

The parameter σ can take on integer values.

Proof. First, we establish Theorem 2 under the more restrictive condition

(2.4) $$\sigma \neq \text{an integer}.$$

Our proof follows from a series of lemmas, but before presenting them, we give a general discussion of the salient features of the problem.

We begin by noticing that condition (i) deletes from $\mathcal{D}_\xi^{\#}$ small discs around the singular points $iq\pi/2$, q an integer, i.e., $|n^*(2\xi-iq\pi)|^{-1}=o(1)<(16)^{-1}$, $n\to\infty$ (see Lemma 1) and that condition (ii), except in the "passageways" $0<|\beta|<\pi/2$, deletes a strip to the left of the imaginary axis, to insure that $|\alpha n^*|\to\infty$, $n\to\infty$, for $\xi=\alpha+i\beta\in\mathcal{D}_\xi$.

In light of Remark 2, and the notation of this theorem, we can write Corollary 1.1 in the form

(2.5)

$$g_n(w)=\sqrt{\pi}\,w^\gamma(1+w)^{-\gamma-\lambda}e^{-2n^*\xi}\frac{\Gamma(n^*+1-\lambda)}{\Gamma(n^*+\lambda)}\frac{\Gamma(-\sigma)}{2\pi i}\int_{\mathcal{C}} e^{2un^*}(2u)^\sigma K(u,\xi)\,d(2u), \qquad \xi\in\mathcal{D}_\xi,$$

(2.6)

$$K(u,\xi)=X^\lambda\left\{\frac{X-1}{2u\coth\xi}\right\}^\sigma \mathcal{S}^*(X)=X^\lambda\left\{\frac{e^{-i\pi}}{2u\coth\xi}\right\}^\sigma \mathcal{S}(X), \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi,$$

$$(2.7)\qquad \mathfrak{S}(X)=(1-X)^{\sigma}\mathfrak{S}^*(X)=\left[e^{i\pi}(X-1)\right]^{\sigma}\mathfrak{S}^*(X),\qquad u\in\mathcal{C},\quad \xi\in\mathfrak{D}_\xi,$$

$$\mathfrak{S}^*(X)=1+\sum_{j=1}^{\infty} d_j\frac{(1-X)^j}{(1+\sigma)_j},\qquad X\in\mathfrak{D},$$

$$X=\left\{\frac{\sinh(\xi-u)}{\sinh\xi}\right\}^{-2},\qquad \frac{X-1}{2u\coth\xi}=\frac{(\sinh u)[\sinh(2\xi-u)]}{(2u\coth\xi)[\sinh(\xi-u)]^2},$$

where the integration contour $\mathcal{C}$ in $\mathfrak{U}$ is described in Corollary 1.1, with $\rho=|9n^*|^{-1}$ in (1.15), see Figs. 2, 5, 6 and note that if $\mathcal{C}$ contains the subcontours $\mathcal{L}_j$, then $|\alpha n^*|>\log|n^*|$. Also, the multiple-valued functions reduce to their principal branches when $u\in\mathcal{C}_0$, $\xi\in\mathfrak{D}_\xi^0$, and are defined elsewhere by analytic continuation with respect to u and ξ. For $u\in\mathfrak{U}$, the functions X, and $(X-1)/(2u\coth\xi)$ are of the form $1+\mathcal{O}(u)$, $u\to 0$.

To apply the contour version of Watson's lemma to (2.5), we need to expand $K(u,\xi)$ in a Maclaurin series at $u=0$, and estimate the behaviour of $K(u,\xi)$ along $\mathcal{C}$. In Lemma 2, it is shown that for $\xi\in\mathfrak{D}_\xi$, $K(u,\xi)$ can be analytically continued into a neighborhood of $u=0$, and that

$$K(u,\xi)=\sum_{k=0}^{\infty}(2u)^k U_k(\xi),\qquad |u|\le 2r,$$

$$=1+\sum_{k=1}^{m-1}(2u)^k U_k(\xi)+(2u)^m K_m(u,\xi),\qquad u\in\mathcal{C},\quad \xi\in\mathfrak{D}_\xi,$$

where r is a number such that $r\ge|8n^*|^{-1}>|9n^*|^{-1}=\rho$, and $U_k(\xi)$ is a polynomial in the two variables $\coth\xi$ and $\coth 2\xi$. The structure of such $U_k(\xi)$ functions is analyzed in Lemma 3. Making use of this expansion for $K(u,\xi)$, together with the fact that

$$\frac{\Gamma(-\sigma)}{2\pi i}\int_{\mathcal{C}} e^{2un^*}(2u)^{\sigma+k}d(2u)=(-1)^k\frac{(\sigma+1)_k}{(n^*)^{\sigma+1+k}},\qquad k \text{ an integer},$$

we can write

$$(2.8)\qquad g_n(w)=\sqrt{\pi}\left[(n^*)^2w\right]^{\gamma}(1+w)^{-\gamma-\lambda}e^{-2n^*\xi}(n^*)^{2\lambda-1}\frac{\Gamma(n^*+1-\lambda)}{\Gamma(n^*+\lambda)}$$

$$\times\left\{1+\sum_{k=1}^{m-1}(-1)^k(\sigma+1)_k U_k(\xi)(n^*)^{-k}+R_m(n^*,\xi)\right\},$$

$$(2.9)\qquad R_m(n^*,\xi)=(n^*)^{1+\sigma}\frac{\Gamma(-\sigma)}{2\pi i}\int_{\mathcal{C}} e^{2un^*}(2u)^{\sigma+m}K_m(u,\xi)d(2u).$$

In Lemma 6, it is shown that there exist nonnegative numbers M,s,t, independent of $\xi=\alpha+i\beta$, n or u, such that for $\xi\in\mathfrak{D}_\xi$,

$$|K_m(u,\xi)|\le\begin{cases} M\omega^m e^{s|u|}\{1+|un^*|\}^t, & u\in\mathcal{C}_0 \text{ or } \mathcal{C}_1,\\ M\omega^m e^{s|\alpha|}|n^*|^t, & u\in\mathcal{L}_1 \text{ or } \mathcal{L}_2,\end{cases}$$

where ω is defined in (2.1). Making the change of variable $v=2un^*$, it is easy to see that the contribution to $R_m(n^*,\xi)$ from $\mathcal{C}_0$ or $\mathcal{C}_1$ is $\mathcal{O}(\omega^m|n^*|^{-m})$, $n\to\infty$. The contributions

to $R_m(n^*,\xi)$ from the $\mathfrak{L}_j$ are much smaller. On $\mathfrak{L}_j^*$, for n^* sufficiently large,

$$\mathrm{Re}(un^*) = -j|\alpha n^* \cos(\arg n^*)|\left\{1+\tau\frac{\tan(\arg n^*)}{j|\alpha|}\right\} < \frac{-|\alpha n^*|}{2},$$

$$\frac{|\alpha|}{2} < |u| < |\alpha n^*|.$$

The same estimates hold on $\mathfrak{L}_j^{**}$, and it is clear that the contributions to $R_m(n^*,\xi)$ from $\mathfrak{L}_j$ are

$$\mathcal{O}\left(\omega^m|n^*|^{\zeta}|\alpha n^*|^{\sigma+m}e^{-|\alpha n^*|+s|\alpha|}\right) = \mathcal{O}\left(\omega^m|n^*|^{-m}\right), \qquad n\to\infty,$$

where $\zeta = 1+t+\mathrm{Re}\,\sigma+\mathrm{Max}\{0,-\mathrm{Re}(\sigma+m)\}$ so that

$$R_m(n^*,\xi) = \mathcal{O}\left(\omega^n|n^*|^{-m}\right), \qquad n\to\infty.$$

Finally, it follows from Lemma 7 that there exist polynomials in λ, $e_{2k}(\lambda)$, such that

$$(n^*)^{2\lambda-1}\frac{\Gamma(n^*+1-\lambda)}{\Gamma(n^*+\lambda)} = 1+\sum_{k=1}^{m-1} e_{2k}(\lambda)(n^*)^{-2k}+(n^*)^{-2m}\mathcal{O}(1),$$

$$n^*\to\infty, \quad |\arg n^*|<\pi.$$

Utilizing this in (2.8) we arrive at the statement of Theorem 2 under the restriction (2.4). Note that

$$S_k(\xi) = (-1)^k\sum_{j=0}^{[k/2]} e_{2j}(\lambda)(\sigma+1)_{k-2j}U_{k-2j}(\xi),$$

has the same structure as the $U_k(\xi)$, so that Lemma 3 is applicable.

We now remove the tentative assumption (2.4), that σ is not an integer. As $g_n(w)$ and the first m terms of (2.8) are well defined whether or not σ is an integer, it is sufficient to show that $R_m(n^*,\xi)$ has an alternate representation which is valid when σ is an integer, and which can be analyzed by the previous methods. Choose m_0 such that $\mathrm{Re}(\sigma+m)>0$ when $m\geq m_0$. As $K_m(u,\xi)$ takes on the same value at corresponding points on $\mathcal{C}$, e.g., $u=xe^{-i\pi}$ and $u=xe^{i\pi}$ on $\mathcal{C}_0$, all factors in the integrand of (2.9) are single-valued, except for $(2u)^{\sigma+m}$. Taking account of the branches of this factor, we can write

(2.10)

$$R_m(n^*,\xi) = (n^*)^{\sigma+1}\frac{(-1)^m}{\Gamma(\sigma+1)}\int_{\mathfrak{L}} e^{-2vn^*}(2v)^{\sigma+m}K_m(ve^{\pm i\pi},\xi)d(2v), \qquad m\geq m_0,$$

where $\mathfrak{L}$ is essentially the straight line $[0,\infty)$ properly deformed to avoid the singularities at $-(k\xi+iq\pi)$, $k=0,1$ or 2, q an integer. Then (2.10) is valid for σ an integer, $\sigma+m_0>0$, although a limiting form of $K_m(ve^{\pm i\pi},\xi)/\Gamma(\sigma+1)$ has to be taken when σ is a negative integer. When $\mathfrak{S}(X)$ is replaced by $\mathfrak{S}(X)/\Gamma(\sigma+1)$, and σ is an integer, Lemmas 1–7 remain valid, so that Theorem 2 is true whether or not σ is an integer, provided $m\geq m_0$. Finally, the restriction on m_0 can be removed, by noticing that

$$S_k(\xi)(n^*)^{-k} = \left(\frac{\omega}{n^*}\right)^k\mathcal{O}(1), \qquad n+\lambda\to\infty, \qquad k=0,1,\cdots,m_0-1.$$

It is left to the reader to show that $\mathcal{D}_\xi$ can be chosen such that $\mathcal{D}_w^* \subset \mathcal{D}_w$. The following lemmas complete the proof of Theorem 2. Unless modified explicitly, the common notation of Theorems 1, Corollary 1.1 and Theorem 2 will be assumed in all the lemmas.

LEMMA 1.

(A) *If* $z=x+iy$, x *and* y *real, then*

$$|x| \le \sinh|x| \le |\sinh z| = \sqrt{\sinh^2 x + \sin^2 y} \le \cosh x,$$

$$\sinh|x| \le |\cosh z| = \sqrt{\sinh^2 x + \cos^2 y} \le \cosh x \le e^{|x|}.$$

(B) *If* x *and* y *are real numbers such that*

$$|y| \le \frac{\pi}{2} \quad \text{and} \quad |\sinh(x+iy)| \le K, \quad \text{then } |x+iy| \le \frac{\pi}{2} K < 2K.$$

(C) *If* $2|z| \le 1$, *then*

$$20\left|\frac{\sinh z}{z} - 1\right| \le 20\left[2\sinh\left(\frac{1}{2}\right) - 1\right] < 1,$$

$$7|\cosh z - 1| \le 1, \qquad 8\left|\frac{\tanh z}{z} - 1\right| \le 1.$$

(D) *For* ξ *general, we have*

$$1 \le 2\omega, \quad |\tanh\xi| \le 2\omega, \quad |\coth\xi| \le 2\omega, \quad |\coth 2\xi| \le 2\omega.$$

Also, if $|\omega/n^*| \le K$, *then* $|\sinh 2\xi| \ge |Kn^*|^{-1}$.

(E) *If* 4δ *is a positive number less than* 1, *then* $|\sinh 2\xi| \ge 4\delta$, *implies*

$$(*) |2\xi + iq\pi| \ge 2\delta, \quad \text{for all integer } q.$$

Conversely, the condition $(*)$ *implies*

$$2|\sinh j\xi| \ge j\delta, \quad 4|\tanh j\xi| \ge j\delta, \quad j=1 \text{ or } 2 \quad \text{and} \quad 3|\coth\xi| \ge \delta.$$

(F) *For* $\xi \in \mathcal{D}_\xi$, *we have*

$$|\omega/n^*| \le \left\{\begin{matrix} \sqrt{2}|n^*|^{-1}, & |w| \ge 2, \\ 3|n^* \sinh 2\xi|^{-1}, & |w| \le 2 \end{matrix}\right\} < \frac{1}{32},$$

$$|n^*(2\xi + iq\pi)| \ge 16 \quad \text{for all integer } q,$$

$$|n^* \sinh j\xi| \ge 4j, \quad |n^* \tanh j\xi| \ge 2j, \quad j=1 \text{ or } 2, \quad \text{and} \quad 3|n^* \coth\xi| \ge 8.$$

Proof. Compute explicitly.

LEMMA 2. *Let*

(1) $16r = \text{Min}\{1, |\tanh\xi|, |\tanh 2\xi|\}$,

(2) $8r^* = \text{Min}\{1, |\tanh\xi|\}$,

(3) $\mathcal{U}_\zeta = \{u: |u| \le \zeta\} \cap \mathcal{U}$,

(4) $Y(u,\xi) = \sinh(\xi - u)/\sinh\xi$, $\arg Y(u,\xi) = 0$, *when* $\xi \in \mathcal{D}_\xi^0$, $u \in [0,\xi) \subset \mathcal{U}$,

(5) $X = X(u,\xi) = \{Y(u,\xi)\}^{-2}$.

Then we have the following results.

(A) *If* $\xi \in \mathcal{D}_\xi$, *then*

$$1 \le 8|n^*|r, \quad \omega^{-1} \le 32r \le 16r^*, \quad \mathcal{U}_{2r} \subset \mathcal{U}_{r^*},$$

and the loop part of $\mathcal{C}_0$, $\mathcal{C}_1$, *i.e.*, $\{u = \rho e^{i\psi}: \rho = |9n^*|^{-1}, -\pi \le \psi \le \pi\}$, *belongs to* $\mathcal{U}_r$.

(B) *For ξ general, we have*

$$Y(u,\xi)=(\cosh u)[1-(\coth\xi)\tanh u],$$

$$\frac{X-1}{2u\coth\xi}=\left(\frac{\sinh u}{u}\right)Y(u,2\xi)X=\left(\frac{\sinh 2u}{2u}\right)[1-(\coth 2\xi)\tanh u]\,X.$$

(C) *If $u\in\mathfrak{U}_{r^*}$, $\xi\in\mathfrak{D}_\xi$, then*

$$|\cosh u-1|<\frac{1}{7},\quad \left|\frac{\sinh 2u}{2u}-1\right|<\frac{1}{20},\quad \left|\frac{\tanh u}{u}-1\right|<\frac{1}{8},$$

$$|(\coth\xi)\tanh u|<\frac{9}{64},\quad |X-1|<\frac{1}{2},\quad |\arg X|<\frac{\pi}{4},$$

and X, X^λ and $\mathfrak{S}_p^*(X)$ are analytic functions of u, with X^λ given by its principal branch. Also, for $t\in\mathfrak{U}_{r^*}$, $\operatorname{Max}|\mathfrak{S}_p^*(X(t,\xi))|$ can be bounded by a constant independent of r, r^* and ξ.

(D) *If $u\in\mathfrak{U}_{2r}$, $\xi\in\mathfrak{D}_\xi$, then*

$$|(\coth 2\xi)\tanh u|<\frac{9}{64},\qquad \left|\arg\left(\frac{X-1}{2u\coth\xi}\right)\right|<\frac{3\pi}{8},$$

and $[(X-1)/(2u\coth\xi)]^\sigma$ is an analytic function of u, given by its principal branch.

(E) *If $\xi\in\mathfrak{D}_\xi$, then $K(u,\xi)$ can be analytically continued as a function of u to $\mathfrak{U}_{2r}$ by the formula*

$$K(u,\xi)=X^\lambda\left\{\frac{X-1}{2u\coth\xi}\right\}^\sigma\mathfrak{S}_p^*(X),\qquad u\in\mathfrak{U}_{2r},\quad \xi\in\mathfrak{D}_\xi. \tag{2.11}$$

In particular, $K(u,\xi)$ then has the Maclaurin expansion

$$K(u,\xi)=\sum_{k=0}^{\infty}(2u)^k U_k(\xi),\qquad u\in\mathfrak{U}_{2r},\quad \xi\in\mathfrak{D}_\xi, \tag{2.12}$$

where the $U_k(\xi)$ have the form

$$U_k(\xi)=\sum_{\substack{s,t\geq 0\\ 0\leq s+t\leq k}} d_{s,t,k}(\coth\xi)^s(\coth 2\xi)^t=(-1)^k U_k(-\xi),$$

the $d_{s,t,k}$ being numbers independent of ξ. The first few $U_k(\xi)$ are

$$U_0(\xi)=1,\qquad (\sigma+1)U_1(\xi)=-S_1(\xi),$$

$$(\sigma+1)_2U_2(\xi)=S_2(\xi)-\frac{(2\lambda-2)_3}{24},$$

$$(\sigma+1)_3U_3(\xi)=-S_3(\xi)+\frac{(2\lambda-2)_3}{24}S_1(\xi),$$

where the $S_k(\xi)$ are defined in Theorem 2. Also, for $t\in\mathfrak{U}_{2r}$, $\operatorname{Max}|K(t,\xi)|$ can be bounded by a constant independent of r, r^* and ξ.

Proof. (A) and (B) follow from direct computations. To show that $(2\omega)(16r)\geq 1$, one makes use of Lemma 1(D). For (C), Lemma 1(C) is applicable, e.g.,

$$\left|r^*\left(\frac{u}{r^*}\right)\left(\frac{\tanh u}{u}\right)\coth\xi\right|<\frac{9}{8}|r^*\coth\xi|\leq\frac{9}{64}.$$

From (B), it is clear that $\log Y(u,\xi)$ is an analytic function of u in $\mathfrak{U}_{r^*}$, which implies $\log X$, X and X^λ are also. Since

$$X-1=(2\coth\xi)(\tanh u)\frac{[1-(\coth 2\xi)\tanh u]}{[1-(\coth\xi)\tanh u]^2},$$

$$2(\coth\xi)(\coth 2\xi)=1+\coth^2\xi,$$

a direct computation shows that $|X-1|\le 1314/3025<\frac{1}{2}$. Since $|z|<1$ implies $|\arg(e^{i0}+z)|\le(\pi/2)|z|$, we have $|\arg X|<\pi/4$, so that X^λ is given by its principal branch, and $\mathfrak{S}_p^*(X)$ is an analytic function of u in $\mathfrak{U}_{r^*}$. Finally, as $u\in\mathfrak{U}_{r^*}$ implies $|X-1|<\frac{1}{2}$, and the coefficients of the series representation for $\mathfrak{S}_p^*(z)$ in (1.3) are independent of r, r^* and ξ, $\operatorname{Max}_{|t|\le 2r}|\mathfrak{S}_p^*(X(t,\xi))|$ can be bounded by a constant independent of r, r^* and ξ. Part (D) is similar. For $u\in\mathfrak{U}_{2r}\subset\mathfrak{U}_{r^*}$, $\log[(X-1)/(2u\coth\xi)]$, $[(X-1)/(2u\coth\xi)]^\sigma$ are analytic functions of u, and

$$|\arg[(X-1)/(2u\coth\xi)]|\le\frac{\pi}{2}\left(\frac{1}{20}+\frac{9}{64}+\frac{1}{2}\right)<\frac{3\pi}{8}$$

implies $[(X-1)/(2u\coth\xi)]^\sigma$ is given by its principal branch. Thus, the function on the right-hand side of (2.11) is an analytic function of u. If in addition, $u\in\mathcal{C}$, $|\arg u|<\pi/2$ and $\xi\in\mathfrak{D}_\xi^0$, so that $|\arg(X-1)|<\pi$, then the right-hand side of (2.11) reduces to $K(u,\xi)=K_p(u,\xi)$. Conversely, the right-hand side of (2.11) serves to analytically continue $K(u,\xi)$ to $\mathfrak{U}_{2r}$ when $\xi\in\mathfrak{D}_\xi$. The rest of (E) follows by explicit computation. The fact that $K(u,\xi)=K(-u,-\xi)$, implies the symmetry property of the $U_k(\xi)$. Finally, as $1<|X|<3$, and $1<4|(X-1)/[2u\coth\xi]|<8$, for $u\in\mathfrak{U}_{2r}$—see the last line of (B), it follows from (2.11) and (C), that $\operatorname{Max}_{|t|\le 2r}|K(t,\xi)|$ can be bounded by a constant independent of r, r^* and ξ.

LEMMA 3. *Let*

$$T_k(\xi)=\sum_{\substack{s,t\ge 0\\ 0\le s+t\le k}} e_{s,t,k}(\coth\xi)^s(\coth 2\xi)^t=(-1)^k T_k(-\xi),$$

where the $e_{s,t,k}$ are numbers independent of ξ. Then:

(A) *There exists a positive constant M_k independent of ξ such that $|T_k(\xi)|\le M_k(2\omega)^k$.*

(B) *There exist constants $a_{j,k}$, $b_{j,k}$ independent of ξ such that*

$$T_k(\xi)=\sum_{j=0}^{[k/2]} a_{k-2j,k}(\coth\xi)^{k-2j}+\sum_{j=0}^{[(k-1)/2]} b_{k-2j,k}(\coth 2\xi)^{k-2j}.$$

Proof. First, from Lemma 1(D), we note that $2\omega\ge 1$, and that

$$|(\coth\xi)^s(\coth 2\xi)^t|\le(2\omega)^{s+t},\qquad s \text{ and } t\ge 0.$$

Then

$$|T_k(\xi)|\le\sum_{j=0}^{k}(j+1)(2\omega)^j\operatorname*{Max}_{s+t=j}|e_{s,t,k}|\le\frac{M_k}{(k+1)}\sum_{j=0}^{k}(2\omega)^j\le M_k(2\omega)^k.$$

Systematically using the relationship

$$2(\coth\xi)^s(\coth 2\xi)^t=(\coth\xi)^{s-1}(\coth 2\xi)^{t-1}+(\coth\xi)^{s+1}(\coth 2\xi)^{t-1},$$

for integers $s, t\ge 1$, we can write

$$T_k(\xi)=\sum_{j=0}^{k}A_{j,k}(\coth\xi)^{k-j}+\sum_{j=0}^{k-1}B_{j,k}(\coth 2\xi)^{k-j},$$

where the $A_{j,k}$, $B_{j,k}$ are numbers independent of ξ. Consider

$$D_k(\xi)=\sum_{\substack{j=0\\ j\text{ odd}}}^{k} A_{j,k}(\coth\xi)^{k-j}+\sum_{\substack{j=0\\ j\text{ odd}}}^{k-1} B_{j,k}(\coth 2\xi)^{k-j}$$

$$=2^{-1}\{T_k(\xi)+(-1)^{k+1}T_k(-\xi)\}\equiv 0.$$

Letting $2\xi\to i\pi$, we see that $D_k(\xi)\equiv 0$ implies $B_{j,k}=0$, j odd. Similarly, letting $\xi\to i\pi$ implies $A_{j,k}=0, j$ odd. Thus, (B) is true with $A_{2j,k}=a_{k-2j,k}$ and $B_{2j,k}=b_{k-2j,k}$.

Remark 5. If $\sqrt{w}=\sinh\xi$ and $\sqrt{1+w}=\coth\xi$, the identities

$$\coth\xi=\frac{2w+2}{\sinh 2\xi},\quad \coth 2\xi=\frac{2w+1}{\sinh 2\xi},\quad 1=\frac{4w(1+w)}{\sinh^2 2\xi}$$

imply that

$$T_k(\xi)=\frac{P_k(w)}{(\sinh 2\xi)^k},$$

where $P_k(w)$ is a polynomial in w of degree k.

LEMMA 4. *Let z belong to a domain $\mathcal{D}_z$ on the Riemann surface $\mathcal{R}_z$, where* $\arg z$, $\arg(1-z)$ *are bounded. Then there exist real numbers M (≥ 0), ν_0, ν_1, ν_∞ independent of z,* $\arg z$ *or* $\arg(1-z)$, *such that uniformly, for $z\in\mathcal{D}_z$,*

(A) $|\mathbb{S}(z)|\leq M|z-1|^{\nu_1}$, *when* $|z-1|\leq\frac{19}{20}$;

(B) $|\mathbb{S}(z)|\leq M|z|^{\nu_0}$, *when* $|z|\leq\frac{19}{20}$;

(C) $|\mathbb{S}(z)|\leq M|z|^{\nu_\infty}$, *when* $|z|\geq\frac{21}{20}$;

(D) $|\mathbb{S}(z)|\leq M$, *when* $\frac{19}{20}\leq|z|\leq\frac{21}{20}$, $|z-1|\geq\frac{19}{20}$.

Proof. Since $\mathbb{S}(z)$ is the solution of a linear, homogeneous, differential equation whose only singularities are the regular singular points $z=0,1$ and ∞, for each choice of integers (s,t) such that

$$|s\pi+\arg z|\leq\pi,\qquad |t\pi+\arg(1-z)|\leq\pi,$$

$\mathbb{S}(z)$ can be expanded in terms of a basis of functions, as a finite sum of algebraic/logarithmic functions in each of the regions described in (A), (B), (C). As logarithmic terms can be dominated by algebraic ones, and only a finite number of (s,t) need be considered, (A), (B) and (C) result. (D) is true, as the region described therein is compact, and contains no singularities of $\mathbb{S}(z)$.

LEMMA 5. *Let $r, r^*, X=X(u,\xi)$ be as in Lemma* 2, *and*

(1) $\Omega=\theta+i\phi=\xi$ *or* 2ξ, $\xi=\alpha+i\beta$, *where θ,ϕ,α,β are real;*

(2) $Y(u,\Omega)=\sinh(\Omega-u)/\sinh\Omega=e^{-u}(1-e^{2u-2\Omega})/(1-e^{-2\Omega})$, $\arg Y(u,\Omega)=0$, *when $\xi\in\mathcal{D}_\xi^0$, $u\in[0,\Omega)\subset\mathcal{U}$;*

(3) $W=Y(u,\Omega)$, X, $(X-1)/(2u\coth\xi)$, *or* $\sinh u/u$, $u\in\mathcal{C}$, $\xi\in\mathcal{D}_\xi$.

Then there exist proper numbers M,s,t (i.e., nonnegative, independent of ξ, n and u), such that for $\operatorname{Re}u\leq -r<0$, $\xi\in\mathcal{D}_\xi$, *we have:*

(A) $|\arg W|\leq M$, $u\in\mathcal{C}$;

(B) $|W|^{\pm 1}\leq Me^{s|u|}\{1+|un^*|\}^t$, $u\in\mathcal{C}_0$ *or* $\mathcal{C}_1$;

(C) $|W|^{\pm 1}\leq Me^{s|\alpha|}|n^*|^t$, $u\in\mathcal{L}_1$ *or* $\mathcal{L}_2$.

Under the stronger conditions $\operatorname{Re}u\leq -r^*$, $\xi\in\mathcal{D}_\xi$, *the results* (A), (B) *and* (C) *are also true for* $W=2u\coth\xi$, $X-1$, *or* $1-X=e^{i\pi}(X-1)$.

Proof. Note that if W_1 and W_2 satisfy (A), (B), (C), then $W_3=W_1\cdot W_2$ also satisfies (A), (B), (C). Thus, for the first part of the lemma when $\operatorname{Re}u\leq -r$, it is sufficient to

prove the lemma for $W=Y(u,\Omega)$ and $(\sinh u)/u$, as the other values of W can be expressed as multiples of these functions—see Lemma 2(B). The contour $\mathcal{C}$ is described in (1.15). Throughout this proof, κ is the parameter occurring in (1.12) and (1.15). In general then, we have the global conditions

$$
\begin{aligned}
&|\operatorname{Im}\Omega|\le 2\kappa, \qquad \xi\in\mathcal{D}_\xi \quad \alpha<0,\\
&|\operatorname{Im}u|<2\kappa+\pi, \qquad u\in\mathcal{C},\\
&|u-\Omega+i\pi q|\ge|9n^*|^{-1}, \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi \quad \text{for all integer } q.
\end{aligned}
\tag{2.13}
$$

Just as in Lemma 1(E), it follows from the last line of (2.13), that

$$
|\sinh(\Omega-u)|>|18n^*|^{-1}, \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi.
\tag{2.14}
$$

The argument computations are straightforward, e.g., if $u\in\mathcal{C}$, $\xi\in\mathcal{D}_\xi$,

$$
|\arg Y(u,\Omega)|\begin{cases}
\le\pi, & u\le -r<0\le\operatorname{Re}\Omega,\\
=\left|\arg\left[e^{\pm i\pi}e^{2\Omega-u}\dfrac{(1-e^{2u-2\Omega})}{(1-e^{2\Omega})}\right]\right|\le 6\kappa+2\pi, & \operatorname{Re}u\le\operatorname{Re}\Omega\le 0,\\
=\left|\arg\left[e^{u}\dfrac{(1-e^{2\Omega-2u})}{(1-e^{2\Omega})}\right]\right|\le 2\kappa+2\pi, & \operatorname{Re}\Omega\le\operatorname{Re}u\le -r,
\end{cases}
$$

and

$$
\left|\arg\left(\frac{\sinh u}{u}\right)\right|=\left|\arg\left[e^{\pm i\pi}e^{-u}\frac{(1-e^{2u})}{2u}\right]\right|\le\kappa+\pi, \qquad \operatorname{Re}u\le -r.
$$

For $u<0$, $u\in\mathcal{C}$, $\xi\in\mathcal{D}_\xi$,

$$
1\le\left|\frac{\sinh u}{u}\right|=|\cosh u|\cdot\left|\frac{\tanh u}{u}\right|\le e^{|u|},
$$

$$
|Y(u,\Omega)|=|\cosh u|\cdot|1-(\coth\Omega)(\tanh u)|\le e^{|u|}\cdot[1+|un^*|],
$$

as $|\coth\Omega|<|n^*|$, by Lemma 1(F). The lower bound for $Y(u,\Omega)$ is more difficult. Equation (2.14) implies

$$
|Y(u,\Omega)|^{-1}\le 18|n^*|\cdot|\sinh\Omega|, \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi,
$$

but unfortunately, in general, this is not sufficiently sharp for our purposes, and we must consider several subcases. As $\Omega=\theta+i\phi$, we have

$$
|Y(u,\Omega)|^{-1}=\sqrt{\frac{\sinh^2\theta+\sin^2\phi}{\sinh^2(\theta-u)+\sin^2\phi}}\le 1, \quad \text{when } 0\le\theta\le\theta-u,
$$

and in particular, for $u\le -r<0<\operatorname{Re}\Omega$. For $\theta\le 0$, we first prove the intermediate result

$$
|Y(u,\Omega)|^{-1}<576e^{2|\theta|}[1+|un^*|], \qquad u\le -r<0.
\tag{2.15}
$$

If $\sin^2\phi\ge\frac{1}{4}$,

$$
|Y(u,\Omega)|^{-1}\le\frac{|\sinh\Omega|}{\sqrt{0+\frac{1}{4}}}\le 2e^{|\theta|}, \qquad u\le -r.
$$

Whereas, if $\sin^2\theta \le \frac{1}{4}$, then $2|\cosh\Omega| \ge 2|\cos\phi| \ge \sqrt{3}$, and it follows from the definition of r, in Lemma 2, that

$$\frac{|\sinh\Omega|}{16r} \le \operatorname{Max}\left\{|\sinh\Omega|, |\cosh\Omega|, 2\left|\cosh\left(\frac{\Omega}{2}\right)\right|^2, \frac{|\cosh 2\Omega|}{2|\cosh\Omega|}\right\} \le 2e^{2|\theta|}.$$

Then (2.14) implies, for $u = -|u| \le -r < 0$,

$$|Y(u,\Omega)|^{-1} \le (18)(16)\left|\frac{\sinh\Omega}{16r}\cdot\frac{r}{u}\cdot un^*\right|$$
$$\le (18)(32)e^{2|\theta|}|un^*| < 576e^{2|\theta|}[1+|un^*|],$$

the desired intermediate result. Equation (2.15) implies (B) when

$$-1 \le \theta \le 0, \qquad u \le -r \quad \text{as } |\theta| \le 1,$$

or

$$\theta \le -1, \qquad u \le \frac{\theta}{2} \quad \text{as } |\theta| \le 2|u|.$$

Finally, if $\theta \le -1$, $\theta/2 \le u \le -r$,

$$|Y(u,\Omega)|^{-1} \le e^{-u}\frac{(1+e^{2\theta})}{(1-e^{2\theta-2u})} < \frac{2e^{|u|}}{1-e^{\theta}} \le \frac{2e^{|u|}}{1-e^{-1}},$$

which completes the proof of (B).

For (C), we note that when $u \in \mathcal{L}_j$, $j = 1$ or 2, then $\alpha < 0$,

$$|\operatorname{Re}(j\alpha - u)| \le |9n^*|^{-1}, \qquad |\operatorname{Re}(k\alpha - u)| \le |\alpha| + |9n^*|^{-1}, \qquad j+k=3.$$

Also, from the special shape of $\mathcal{D}_\xi$ (cf. Theorem 2(ii)), it follows for $\xi \in \mathcal{D}_\xi$, that $|\alpha n^*| \ge 2$. We then have for $u \in \mathcal{L}_j$, $\xi \in \mathcal{D}_\xi$,

$$|Y(u,\Omega)|^{-1} \le \frac{\cosh\operatorname{Re}(\Omega-u)}{\sinh|\theta|} \le \frac{e^{|\operatorname{Re}(\Omega-u)|}|n^*|}{|\alpha n^*|} \le e^{|\alpha|}|n^*|,$$

$$|Y(u,\Omega)|^{-1} \le 18|n^*\cosh\theta| \le 18e^{2|\alpha|}|n^*|,$$

$$\left|\frac{\sinh u}{u}\right| \le \frac{\cosh\operatorname{Re}u}{|\operatorname{Re}u|} \le \frac{e^{|\operatorname{Re}u|}}{j|\alpha|-|9n^*|^{-1}} < \frac{e^{|\operatorname{Re}u|}|n^*|}{|\alpha n^*|-1} \le 2e^{2|\alpha|}|n^*|,$$

$$\left|\frac{\sinh u}{u}\right|^{-1} \le \frac{|u|}{\sinh|\operatorname{Re}u|} \le 1 + \frac{|\operatorname{Im}u|}{|\operatorname{Re}u|} < 1 + \frac{(2\kappa+\pi)|n^*|}{|\alpha n^*|-1} < 2(\kappa+\pi)|n^*|,$$

which completes the lemma, when $\operatorname{Re}u \le -r$.

Similarly for $\operatorname{Re}u \le -r^* < -r < 0$, $\xi \in \mathcal{D}_\xi$, consider $W = 2u\coth\xi$. Under these conditions, we have

$$|\arg(2u\coth\xi)| \le |\arg u| + \left|\arg\left(\frac{1+e^{-2\xi}}{1-e^{-2\xi}}\right)\right| < 3\pi, \qquad \operatorname{Re}\xi \ge 0, \quad u \in \mathcal{C},$$

$$\le \left|\arg\left(e^{\pm i\pi}u\frac{(1+e^{2\xi})}{(1-e^{2\xi})}\right)\right| < 4\pi, \qquad \operatorname{Re}\xi \le 0, \, u \in \mathcal{C},$$

$$|u\coth\xi|\le|un^*|<1+|un^*|, \qquad u\in\mathcal{C}_0 \text{ or } \mathcal{C}_1,$$
$$\le[|\mathrm{Re}\,u|+|\mathrm{Im}\,u|]|n^*|\le[1+2\kappa+\pi+2|\alpha|]|n^*|$$
$$\le e^{2\kappa+\pi+2|\alpha|}|n^*|, \qquad u\in\mathcal{L}_1 \text{ or } \mathcal{L}_2,$$

$$|u\coth\xi|^{-1}\le\frac{|\tanh\xi|}{r^*}\le\mathrm{Max}\{8,64|u\tanh\xi|\}\le 64[1+|un^*|], \qquad u\in\mathcal{C}_0 \text{ or } \mathcal{C}_1,$$

$$|u\coth\xi|^{-1}\le\frac{|n^*|}{|\mathrm{Re}\,u|}\le\frac{|n^*|^2}{|\alpha n^*|-1}\le|n^*|^2, \qquad u\in\mathcal{L}_1 \text{ or } \mathcal{L}_2,$$

where again the special shape of $\mathcal{D}_\xi$ has been used. Thus, $2u\coth\xi$, $(2u\coth\xi)[(X-1)/(2u\coth\xi)]=X-1$ and $1-X=e^{i\pi}(X-1)$, satisfy (A), (B) and (C), completing the lemma.

LEMMA 6. *Let $U_k(\xi)$ be as in Lemma* 2(E), *and*

(1) $m+1$ *be a positive integer*;

(2) $K_m(u,\xi)$ *for $u\in\mathcal{C}$, $\xi\in\mathcal{D}_\xi$ be defined by*

$$K(u,\xi)=\sum_{k=0}^{m-1}(2u)^kU_k(\xi)+(2u)^mK_m(u,\xi), \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi.$$

Then there exist proper numbers $M(\ge1)$, s,t (i.e. nonnegative, independent of ξ, n and u), such that for $u\in\mathcal{C}$, $\xi\in\mathcal{D}_\xi$, we have:

(A) $|K_m(u,\xi)|\le M\omega^m e^{s|u|}\{1+|un^*|\}^t$, $u\in\mathcal{C}_0$ *or* $\mathcal{C}_1$;

(B) $|K_m(u,\xi)|\le M\omega^m e^{s|\alpha|}|n^*|^t$, $u\in\mathcal{L}_1$ *or* $\mathcal{L}_2$.

The constant M depends on m.

Proof. In what follows, we will use the notation of Lemma 2. From Lemma 2(E) and Cauchy's estimate for the remainder of a series, it follows that for $|u|\le r$, $\xi\in\mathcal{D}_\xi$,

$$|(2u)^mK_m(u,\xi)|\le\frac{|2u|^m}{(3r)^m[1-|2u|/(3r)]}\operatorname*{Max}_{2|t|=3r}|K(t,\xi)|$$
$$\le\frac{3|64u\omega/3|^m}{(32r\omega)^m}\operatorname*{Max}_{2|t|=3r}|K(t,\xi)|\le|2u\omega|^mM,$$

for a proper constant M, which implies (A) and (B) when $|u|\le r$. Here we have used the facts, that $1\le32r\omega$ (Lemma 2(A)), that $K(u,\xi)$ can be analytically continued as a function of u to $\mathcal{U}_{2r}$ (see (2.12)), and that $\mathrm{Max}_{2|t|=3r}|K(t,\xi)|$ can be bounded by a constant independent of r and ξ (see Lemma 2(E)).

For $|u|\ge r$, we have $|32u\omega|\ge|32r\omega|\ge1$, and Lemma 3 implies that

(2.16)
$$|(2u)^mK_m(u,\xi)|=|K(u,\xi)-\sum_{k=0}^{m-1}(2u)^kU_k(\xi)|\le|K(u,\xi)|+\sum_{k=0}^{m-1}|2u|^kM_k(2\omega)^k$$
$$\le|32u\omega|^m\left[|K(u,\xi)|+\sum_{k=0}^{m-1}M_k8^{-k}\right]$$
$$\le|32u\omega|^mM\,\mathrm{Max}\{1,|K(u,\xi)|\}, \qquad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_\xi,$$

for a proper constant $M(\ge1)$. Once the growth properties of $K(u,\xi)$, as defined by (2.11), are established for $\mathrm{Re}\,u\le-r$, $u\in\mathcal{C}$ and $\xi\in\mathcal{D}_\xi$, the remaining cases of the lemma will follow directly from (2.16).

For $|u|\leq r^*$, Lemma 2(C) implies that $|X-1|<1/2$, that $\mathfrak{S}^*(X)$ is bounded by a proper number M, and hence it follows from (2.11), that we have

$$|K(u,\xi)|\leq M\left|X^{\lambda}\left(\frac{X-1}{2u\coth\xi}\right)^{\sigma}\right|, \qquad |u|\leq r^*, \quad u\in\mathcal{C}, \quad \xi\in\mathcal{D}_{\xi}.$$

Lemma 5 then implies (A) and (B), when $r\leq|u|\leq r^*$, $u\in\mathcal{C}$, $\xi\in\mathcal{D}_{\xi}$. For $|u|\geq r^*$, the behaviour of $K(u,\xi)$ is essentially determined by whether

$$X=X(u,\xi)=\{Y(u,\xi)\}^{-2}=\left\{\frac{\sinh(\xi-u)}{\sinh\xi}\right\}^{-2},$$

is near 0, 1 or ∞, e.g., if q is an integer, and

$u=$	0	$\xi+i\pi q$	$2\xi+i\pi q$	$-\infty$
then $Y(u,\xi)=$	1	0	$(-1)^{q+1}$	∞
and $X(u,\xi)=$	1	∞	1	0

As u progresses along $\mathcal{C}$, X varies among regions of the form described in Lemma 4, (A)–(D). Since Lemma 5 implies $\arg X$ and $\arg(1-X)$ are bounded for $|u|\geq r^*$, $u\in\mathcal{C}$, $\xi\in\mathcal{D}_{\xi}$, only a finite number of such regions are actually entered. In each of these regions, Lemma 4 can be used to estimate $\mathfrak{S}(X)$ in terms of X and $X-1$. Combining these estimates with (2.16) and the results of Lemma 5 for $\operatorname{Re}u\leq -r^*$, we are lead directly to (A) and (B), which completes the lemma.

LEMMA 7. *For λ bounded,*

$$(n+\lambda)^{2\lambda-1}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}=1+\sum_{k=1}^{m-1}e_{2k}(\lambda)(n+\lambda)^{-2k}+\mathcal{O}\big((n+\lambda)^{-2m}\big),$$

$$n+\lambda\to\infty, \qquad |\arg(n+\lambda)|<\pi,$$

$$e_{2k}(\lambda)=\frac{(2\lambda-1)_{2k}}{(2k)!}B_{2k}^{(2-2\lambda)}(1-\lambda),$$

where $B_j^{(\sigma)}(x)$ is the generalized Bernoulli polynomial defined by

$$\left(\frac{t}{e^t-1}\right)^{\sigma}e^{xt}=\sum_{j=0}^{\infty}\frac{t^j}{j!}B_j^{(\sigma)}(x), \qquad |t|<2\pi.$$

In particular,

$$e_2(\lambda)=\frac{(2\lambda-2)_3}{24}, \qquad e_4(\lambda)=\frac{(2\lambda-2)_5(5\lambda-6)}{2880}.$$

Proof. See [3].

This completes the proof of Theorem 2.

COROLLARY 2.1. *With the same conditions and notation as Theorem* 2,

$$g_n(w)=\sqrt{\pi}\left[(n+\lambda)^2w\right]^{\gamma}(1+w)^{-\gamma-\lambda}\exp\{-2(n+\lambda)\xi+T(n+\lambda,\xi)\},$$

$$T(n+\lambda,\xi)=\sum_{k=1}^{m-1}T_k(\xi)(n^*)^{-k}+\left(\frac{\omega}{n^*}\right)^m\mathcal{O}(1), \qquad n+\lambda\to\infty,$$

$$T_k(\xi)=\sum_{j=0}^{[k/2]}U_{k-2j,k}(\coth\xi)^{k-2j}+\sum_{j=0}^{[(k-1)/2]}V_{k-2j,k}(\coth 2\xi)^{k-2j},$$

$$U_{11}=d_1-(\sigma+1)(\sigma+\lambda),$$
$$2U_{22}=2d_2-[d_1+(3\sigma+2\lambda+3)]d_1+(\sigma+1)(\sigma+\lambda)(\sigma+\lambda+1),$$
$$6U_{02}=3(\sigma+1)d_1-(\sigma+1)[2\sigma^2+\sigma(3\lambda+1)+3\lambda]+\lambda(\lambda-1)(2\lambda-1),$$
$$\begin{aligned}24U_{33}=&24d_3+8d_1^3-24d_1d_2-12[6\sigma+(4\lambda+9)]d_2\\&+12[3\sigma+(2\lambda+3)]d_1^2+3[21\sigma^2+\sigma(28\lambda+45)+8(\lambda+1)(\lambda+3)]d_1\\&-(\sigma+1)(\sigma+\lambda)[13\sigma^2+\sigma(12\lambda+25)+4(2\lambda^2+6\lambda+3)],\end{aligned}$$
$$\begin{aligned}8U_{13}=&4(2\sigma+3)d_2-4(\sigma+1)d_1^2-[15\sigma^2+\sigma(12\lambda+31)+16(\lambda+1)]d_1\\&+(\sigma+1)(\sigma+\lambda)[5\sigma^2+\sigma(12\lambda+11)+4(2\lambda+1)],\end{aligned}$$
$$2V_{11}=-4V_{22}=\sigma(\sigma+1),\qquad 24V_{33}=-\sigma(\sigma+1)(\sigma+3)(\sigma-2),$$
$$8V_{13}=\sigma(\sigma+1)\{2d_1-[\sigma^2+\sigma(2\lambda+1)+2(\lambda+1)]\},\ \textit{etc.}$$

In particular, $T_1(\xi)=S_1(\xi)$.

Proof. $T(n+\lambda,\xi)=\log S(n+\lambda,\xi)$, and the rest follows by explicit computation.

COROLLARY 2.2. *If in Theorem* 2, *or Corollary* 2.1, *one uses the large variable* N,

$$N=\sqrt{n(n+2\lambda)},\quad \textit{or}\quad (n+\lambda)=\sqrt{N^2+\lambda^2},$$

instead of $(n+\lambda)$, *one obtains asymptotic expansions of the same general form for* $e^{2(n+\lambda)\xi}g_n(w)$, *but with* $(n+\lambda)$ *replaced by* N, *and* $S_k(\xi)$, $T_k(\xi)$ *replaced by* $S_k^*(\xi)$, $T_k^*(\xi)$, *respectively. In particular*

$$[N^2w]^{-\gamma}(1+w)^{\gamma+\lambda}e^{2(n+\lambda)\xi}g_n(w)=\sqrt{\pi}\,S^*(N,\xi)=\sqrt{\pi}\,e^{T^*(N,\xi)},$$
$$S^*(N,\xi)=1+\sum_{k=1}^{m-1}S_k^*(\xi)N^{-k}+\left(\frac{\omega}{N}\right)^m\mathcal{O}(1),\qquad N\to\infty,$$
$$T^*(N,\xi)=\sum_{k=1}^{m-1}T_k^*(\xi)N^{-k}+\left(\frac{\omega}{N}\right)^m\mathcal{O}(1),\qquad N\to\infty,$$

and

$$S_1^*(\xi)=S_1(\xi)=T_1^*(\xi)=T_1(\xi),$$
$$S_2^*(\xi)=S_2(\xi)+\gamma\lambda^2,\qquad T_2^*(\xi)=T_2(\xi)+\gamma\lambda^2,$$
$$S_3^*(\xi)=S_3(\xi)+\left(\frac{\lambda^2}{2}\right)(2\gamma-1)S_1(\xi),\qquad T_3^*(\xi)=T_3(\xi)+\left(\frac{-\lambda^2}{2}\right)T_1(\xi).$$

Proof. Note that

$$(n+\lambda)^{-\nu}=N^{-\nu}\,{}_1F_0\left(\frac{\nu}{2}\;\middle|\;\frac{-\lambda^2}{N^2}\right).$$

The rest follows by computation.

The asymptotic expansion of the $l_{n,j}(w)$ comes from the following general result.

PROPOSITION 1. *Let*

$$I_n(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_0^\infty e^{-2v(n+\lambda)}(\sinh y)^{-2\lambda}X^\sigma(\log X)^kE(v,X)\,dv,$$

where $X=w(\sinh v)^{-2}$. *Assume that*:

(1) n *is a large parameter such that* $|n|\to+\infty$, $\arg n=\mathcal{O}(n^{-1})$ *as* $n\to\infty$;

(2) λ,σ *are complex parameters, bounded with respect to* n, *such that* $2\mathrm{Re}(\lambda+\sigma)<1$;

(3) w *is a complex parameter such that* $|\arg w|\le 2\pi-6\varepsilon$, ε *a small positive number independent of* n, *and* $|\Omega|\ge 5$, $\Omega=w(n+\lambda)^2$;

(4) k *is a integer* ≥ 0;

(5) $E(v,X)$ *is an analytic and uniformly bounded function of* v *for* $|\arg X|\le\pi-2\varepsilon$. *Then* $I_n(w)=\mathcal{O}(\Omega^{\sigma}[\log|\Omega|]^k)$, $n+\lambda\to\infty$.

Proof. For $|\arg w|\le 2\pi-6\varepsilon$, choose the real number ϕ such that $2|\phi|\le\pi-4\varepsilon$, and $|\arg w-2\phi|\le\pi-2\varepsilon$. Then the integration contour $\mathcal{C}_0$: $0\le v<\infty$, can be deformed into

$$\mathcal{C}_\phi\colon v=\log\left(t+\sqrt{1+t^2}\right),\qquad t=\sinh v=ue^{i\phi},\qquad 0\le u<\infty,\quad 2|\phi|<\pi.$$

This contour is shown for $0<2\phi<\pi$ in Fig. 7. As v varies along $\mathcal{C}_\phi$ between 0 and ∞, $|v|$ is increasing, $|\arg v|$ is nonincreasing,

$$2|\arg v|\le 2|\phi|\le\pi-4\varepsilon,\qquad |\arg X|=|\arg w-2\phi|\le\pi-2\varepsilon,$$

and $E(v,X)=\mathcal{O}(1)$, uniformly for $v\in\mathcal{C}_\phi$. Let $\mathcal{C}_\phi$ cross $|v|=1$ at $v=e^{i\theta}$. Making the change of variable $t=\sinh v$, we can write

$$I_n(w)=w^{\sigma}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_0^{\infty e^{i\phi}}e^{-2v(n+\lambda)}t^{-2\lambda-2\sigma}(\log X)^k\frac{E(v,X)}{\sqrt{1+t^2}}\,dt.$$

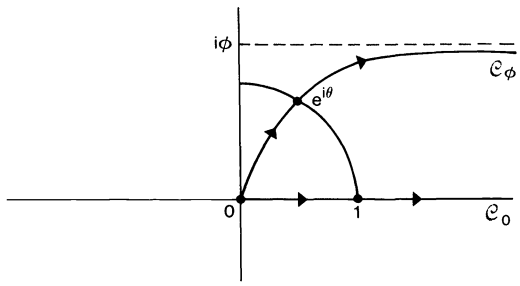


FIG 7. v-*plane*.

The following approximations are valid for $t=\sinh v=ue^{i\phi}$, $u\ge 0$. For $|v|\le 1$,

$$5|tv^{-1}-1|\le 5[(\sinh 1)-1]<1,$$

which implies

$$4\le 5|tv^{-1}|\le 6,\quad 2|\arg(tv^{-1})|<\pi,$$

whereas, for $|v|\ge 1$,

$$|2te^{-v}-1|=|e^{-2v}|\le e^{-2|v|\cos\phi}\le e^{-2\sin 2\varepsilon}<1,$$

which implies that

$$e^{-\sin 2\varepsilon}\sin 2\varepsilon\le|te^{-v}|<1,\qquad 2|\arg(te^{-v})|<\pi.$$

Also,

$$\begin{aligned}|\log X|^k&\le\left[|\arg X|+\log|\Omega|+2|\log|u(n+\lambda)||\right]^k\\&\le[3\log|\Omega|]^k\left[1+|\log|u(n+\lambda)||\right]^k,\end{aligned}$$

and

$$|1+t^2|=\sqrt{(1-u^2)^2+4u^2\cos^2\phi}$$
$$\geq 1-u^2\geq\frac{1}{2}, \qquad 0\leq u\leq 2^{-1/2},$$
$$\geq 2u\cos\phi\geq\sqrt{2}\sin 2\varepsilon, \qquad 2^{-1/2}\leq u,$$
$$\geq \operatorname{Min}\left\{\frac{1}{2},\sqrt{2}\sin 2\varepsilon\right\}=\sqrt{2}\sin 2\varepsilon,$$

for ε sufficiently small.

Using these estimates, we have

(2.17)

$$\Omega^{-\sigma}[\log|\Omega|]^{-k}I_n(w)$$
$$=\mathcal{O}\left((n+\lambda)^{1-2\lambda-2\sigma}\int_0^\infty|e^{-2v(n+\lambda)}|u^{-2\operatorname{Re}(\lambda+\sigma)}[1+|\log|u(n+\lambda)||]^k du\right).$$

We will break this contour at

$$u_1=|\sinh(e^{i\theta})|=\sqrt{\sinh^2(\cos\theta)+\sin^2(\sin\theta)}\geq\sqrt{\cos^2\theta+4\pi^{-2}\sin^2\theta}\geq 2\pi^{-1}.$$

Let $\rho=\sin\varepsilon>0$. On $[0,u_1]$,

$$|e^{-2v(n+\lambda)}|\leq e^{-2\rho|v(n+\lambda)|}\leq e^{-(5/3)\rho|u(n+\lambda)|}\leq e^{-\rho|u(n+\lambda)|},$$

and on $[u_1,\infty)$, $|u(n+\lambda)|\geq 1$ for n sufficiently large, and

$$|e^{-v(n+\lambda)}|\leq e^{-\rho|n+\lambda|}<1, \qquad u_1\leq u,$$
$$\leq|(te^{-v})^{(n+\lambda)}t^{-(n+\lambda)}|\leq e^{(\pi/2)|\operatorname{Im}(n+\lambda)|}|t^{-(n+\lambda)}|, \qquad 1\leq u.$$

Thus, if I_n^* denotes the argument of the $\mathcal{O}$ symbol in (2.17), we have

$$I_n^*=\mathcal{O}\left(\int_0^{u_1|n+\lambda|}e^{-\rho x}x^{-2\operatorname{Re}(\lambda+\sigma)}[1+|\log|x||]^k dx\right)$$
$$+\mathcal{O}\left((n+\lambda)^{k+1-2\lambda-2\sigma}e^{-\rho|n+\lambda|}\int_{u_1}^\infty|e^{-v(n+\lambda)}|u^{k-2\operatorname{Re}(\lambda+\sigma)}du\right)$$
$$=\mathcal{O}(1)+\mathcal{O}\left(\int_{u_1}^{\operatorname{Max}(1,u_1)}u^{k-2\operatorname{Re}(\lambda+\sigma)}du\right)+\mathcal{O}\left(\int_1^\infty u^{k-\operatorname{Re}(n+3\lambda+2\sigma)}du\right)$$
$$=\mathcal{O}(1) \quad \text{as } n+\lambda\to\infty,$$

which proves the proposition.

Remark 6. Proposition 1 is applicable to the integral

$$\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_0^\infty e^{-2v(n+\lambda)}(\sinh v)^{-2\lambda}X^\sigma(\log X)^k dv$$
$$=w^\sigma\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\sum_{r=0}^k\binom{k}{r}(-2)^r(\log w)^{k-r}F_r(2n+2\lambda+1,1-2\lambda-2\sigma),$$

where the notation and conditions are the same as in Proposition 1, and

$$(2.18)\quad F_r(x,y)=\int_0^\infty e^{-u(x-1)}(\sinh u)^{y-1}[\log(\sinh u)]^r du, \qquad \operatorname{Re}x>\operatorname{Re}y>0,$$

$$=\frac{\partial^r}{\partial y^r}\left\{2^{-y}\frac{\Gamma\left(\frac{x-y}{2}\right)\Gamma(y)}{\Gamma\left(\frac{x+y}{2}\right)}\right\}.$$

THEOREM 3. *Let $p+1$, j be positive integers with $j\le p$, n be a large parameter such that $|n|\to+\infty$, $\arg n=\mathcal{O}(n^{-1})$ as $n\to\infty$, and λ, a_k, b_k be complex parameters such that:*

(1) *they are bounded with respect to n;*

(2) *$b_k-a_j\neq$ a negative integer, $k=1,\cdots,p+2$.*

Then, for $[w(n+\lambda)^2]^{-1}=o(1)$ as $n\to\infty$, and $m+1$ an arbitrary positive integer,

$$l_{n,j}(w)=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G^{p+3,2}_{p+3,p+3}\left(w\,\middle|\,\begin{matrix}1-n-2\lambda,a_j,a_P,n+1\\ b_{P+2},a_j\end{matrix}\right)$$

$$=\sum_{k=0}^{m-1}(-1)^k C_{k,j}\frac{(n+2\lambda)_{-k-1+a_j}}{(n+1)_{k+1-a_j}}w^{-k-1+a_j}+\mathcal{O}\left(\left[w(n+\lambda)^2\right]^{-m-1-a_j}\right)$$

$$\sim C_{0,j}\frac{(n+2\lambda)_{-1+a_j}}{(n+1)_{1-a_j}}w^{-1+a_j}{}_{p+3}F_{p+2}\left(\begin{matrix}1,1+b_{P+2}-a_j\\ 1+a_P-a_j,n+2-a_j,-n-2\lambda+2-a_j\end{matrix}\,\middle|\,\frac{1}{w}\right),$$

$$n+\lambda\to\infty, \qquad |\arg w|\le 2\pi-6\varepsilon,$$

where ε is a small positive number independent of n, and

$$C_{k,j}=\frac{\Gamma(k+1+b_{P+2}-a_j)}{\Gamma(k+1+a_P-a_j)}.$$

Proof. It follows from (2.18) with $r=0$ and the Mellin–Barnes integral representation for $l_{n,j}(w)$, that for n sufficiently large,

$$(2.19)\qquad l_{n,j}(w)=2\sqrt{\pi}\,\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_0^\infty e^{-2v(n+\lambda)}(\sinh v)^{-2\lambda}G_j(X)\,dv,$$

$$2\operatorname{Re}(\lambda+a_j)<3,\ 0<w,\ X=w(\sinh v)^{-2},$$

$$G_j(X)=G^{p+3,1}_{p+3,p+3}\left(X\,\middle|\,\begin{matrix}a_j,a_P,\frac{1}{2}-\lambda,1-\lambda\\ b_{P+2},a_j\end{matrix}\right).$$

Since

$$G_j(X)=\sum_{k=0}^{m-1}(-1)^k\frac{C_{k,j}X^{-k-1+a_j}}{\Gamma\left(k-a_j+\frac{3}{2}-\lambda\right)\Gamma(k-a_j+2-\lambda)}+G_{j,m}(X),$$

$$G_{j,m}(X)=(-1)^m G^{p+3,1}_{p+3,p+3}\left(X\,\middle|\,\begin{matrix}-m+a_j,a_P,\frac{1}{2}-\lambda,1-\lambda\\ -m+a_j,b_{P+2}\end{matrix}\right),$$

we can rewrite (2.19) in the form

$$l_{n,j}(w)=\sum_{k=0}^{m-1}(-1)^k C_{k,j}\frac{(n+2\lambda)_{-k-1+a_j}}{(n+1)_{k+1-a_j}}w^{-k-1+a_j}+R_{m,j}(n,w), \tag{2.20}$$

$$R_{m,j}(n,w)=2\sqrt{\pi}\,\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\int_0^\infty e^{-2v(n+\lambda)}(\sinh v)^{-2\lambda}\ G_{j,m}(X)\,dv,$$

$$2\operatorname{Re}(\lambda+a_j)<2m+3,\quad 0<w.$$

Note that for v near 0, X is large, and that when $|X|>1$, $|\arg X|\le\pi-2\varepsilon$,

$$G_{j,m}(X)=(-1)^m\frac{C_{m,j}X^{-m-1+a_j}}{\Gamma\left(m+\frac{3}{2}-\lambda-a_j\right)\Gamma(m+2-\lambda-a_j)}$$

$$\times{}_{p+3}F_{p+2}\left(\begin{matrix}1,m+1+b_{P+2}-a_j\\ m+1+a_P-a_j,m+\frac{3}{2}-\lambda-a_j,m+2-\lambda-a_j\end{matrix}\middle|\frac{-1}{X}\right).$$

Alternately, when $|X|<1$, $|\arg X|\le\pi-2\varepsilon$, it follows directly from the Mellin–Barnes integral for $G_{j,m}(X)$, that $G_{j,m}(X)$ has the convergent series representation,

$$G_{j,m}(X)=X^{-m+a_j}F_{0,j,m}(X)+\sum_{k=1}^{p+2}X^{b_k}F_{k,j,m}(X),\qquad 0<|X|<1,$$

where $F_{0,j,m}(X)$ is a hypergeometric function of the form ${}_{p+3}F_{p+2}(-X)$, and the $F_{k,j,m}(X)$ are related series, possibly involving positive integer powers, of $\log X$ as multiplicative factors. For m sufficiently large, say $m\ge m_0$, $X^{-m+a_j}F_{0,j,m}(-X)$ is the dominant term of this series representation as $X\to 0$. Finally, as the only singular points of $G_{j,m}(X)$ are at $X=0$, -1 and ∞, we have

$$G_{j,m}(X)=\mathcal{O}(X^{-m-1+a_j})\quad\text{uniformly for } |\arg X|\le\pi-2\varepsilon.$$

By analytic continuation with respect to w, we see that (2.20) is actually valid for $|\arg w|\le 2\pi-6\varepsilon$. When Proposition 1 with $k=0$, is applied to $R_{m,j}(n,w)$, we get

$$R_{m,j}(n,w)=\mathcal{O}\left(\left[w(n+\lambda)^2\right]^{-m-1+a_j}\right),\qquad n+\lambda\to\infty,\quad m\ge m_0. \tag{2.21}$$

As each of the m terms in (2.20) can be estimated explicitly, the tentative assumption $m\ge m_0$ in (2.21) can be dropped, which completes the theorem.

Remark 7. Let $l_{n,j}^{\langle m\rangle}(w)$ denote the first m terms of the asymptotic series for $l_{n,j}(w)$ as given in Theorem 3. Under the conditions of Theorem 3, we then have

$$l_{n,j}(w)=l_{n,j}^{\langle m\rangle}(w)+\mathcal{R}_{m,j}(w),$$

$$\mathcal{R}_{m,j}(w)=\mathcal{O}\left(\left[w(n+\lambda)^2\right]^{-m-1-a_j}\right),\qquad n+\lambda\to\infty,\quad |\arg w|<2\pi.$$

Consider the function

$$\mathcal{L}_n(w)=\sum_{j=1}^{p}A_j l_{n,j}(w)=\mathcal{L}_n^{\langle m\rangle}(w)+\mathcal{R}_m(w),$$

$$\mathcal{L}_n^{\langle m\rangle}(w)=\sum_{j=1}^{p}A_j l_{n,j}^{\langle m\rangle}(w),\qquad \mathcal{R}_m(w)=\sum_{j=1}^{p}A_j\mathcal{R}_{m,j}(w),$$

where the A_j are well-defined constants depending on the parameters a_k, b_k. Then the asymptotic expansion of $\mathfrak{L}_n(w)$ follows directly from Theorem 3, i.e.,

$$\mathfrak{L}_n(w)=\mathfrak{L}_n^{\langle m\rangle}(w)+\sum_{j=1}^{p}A_j\mathcal{O}\Big(\big[w(n+\lambda)^2\big]^{-m-1-a_j}\Big),$$

$$n+\lambda\to\infty,\quad |\arg w|<2\pi.$$

For particular limiting configurations of the parameters a_k and b_k, it may happen that one or more of the A_j become singular, and in such cases, $\mathfrak{L}_n(w)$ need not have an asymptotic expansion, in general. However, in special cases such as

$$A_j=\frac{\Gamma(1+a_P-a_j)}{\Gamma(1+b_{P+2}-a_j)},$$

an asymptotic expansion of $\mathfrak{L}_n(w)$ can be found, by taking appropriate limits in the $\mathfrak{L}_n^{\langle m\rangle}(w)$, and modifying the proof of Theorem 3 to show that $\mathfrak{R}_m(w)$ has an appropriate order estimate. The crucial steps of this modified proof would depend on showing that for appropriate parameters σ and k,

$$\sum_{j=1}^{p}A_jG_j(X)=\mathcal{O}\big(X^{\sigma}(\log X)^k\big)\quad\text{uniformly for } |\arg X|\le\pi-2\varepsilon,$$

and would make use of Proposition 1 in its full generality. Typically, this limit procedure introduces nonnegative integer powers of $\log[w(n+\lambda)^2]$ into $\mathfrak{L}_n^{\langle m\rangle}(w)$.

Theorems 2 and 3 can be extended by making use of the following analytic continuation results for the principal branches of $g_n(w)$ and the $l_{n,j}(w)$.

THEOREM 4. *For $p+1$ a positive integer, $j,q=1,\cdots,p$ and $k=1,\cdots,p+2$, let*

(1) $a_j-a_q\neq$ *an integer* $(j\neq q)$;

(2) $b_k-a_j\neq$ *a negative integer*;

(3) $n+2\lambda-1+b_k\neq$ *a negative integer*;

(4) $\sigma=-2\lambda-2\gamma=\frac{1}{2}-2\lambda+\sum_{j=1}^{p}a_j-\sum_{k=1}^{p+2}b_k$.

Then the principal branches of $g_n(w)$ and $l_{n,j}(w)$ have the following analytic continuations. When $|w|>1$,

(A) $g_n(we^{i2\pi})=e^{-i2\pi(n+2\lambda)}g_n(w)$;

(B) $l_{n,j}(we^{i2\pi})=e^{i2\pi a_j}l_{n,j}(w)+(-i2\pi)e^{i\pi a_j}g_n(we^{i\pi})$.

When $0<|w|<1$,

(C) $g_n(we^{i2\pi})=\sum_{j=1}^{p}C_jl_{n,j}(we^{i\pi})+D_1g_n(we^{-i2\pi})+D_2g_n(w)$;

(D) $l_{n,j}(we^{i2\pi})=e^{i2\pi a_j}l_{n,j}(w)+(-i2\pi)e^{i\pi a_j}g_n(we^{i\pi})$;

where

$$D_1=e^{i4\pi\gamma},\qquad D_2=e^{i4\pi\gamma}\left\{\sum_{j=1}^{p}e^{-i2\pi a_j}-\sum_{k=1}^{p+2}e^{-i2\pi b_k}\right\},$$

$$C_j=(2\pi)^2e^{i\pi(1/2+2\gamma-a_j)}\frac{\Gamma\big(a_P^{*j}-a_j\big)\Gamma(1-a_P+a_j)}{\Gamma(b_{P+2}-a_j)\Gamma(1-b_{P+2}+a_j)}.$$

Formulae (A)–(D) *remain true if* (i) *is replaced by* $(-i)$ *and the constants C_j, D_k are replaced by their complex conjugates $\overline{C}_j$, $\overline{D}_k$.*

In $|w-1|<1$, the principal branch of $l_{n,j}(w)$ is analytic, so that $w=1$ is only an apparent singularity. The principal branch of $g_n(w)$ has the following analytic continuations. For $|w|>1$, we have

(E)

$$e^{\pm i\pi(n+2\lambda)}g_n(we^{\pm i\pi})$$
$$=\frac{\Gamma(n+1)\Gamma(n+2\lambda+b_{P+2})w^{-n-2\lambda}}{\Gamma(n+2\lambda)\Gamma(2n+2\lambda+1)\Gamma(n+2\lambda+a_P)}\,{}_{p+2}F_{p+1}\left(\begin{matrix}n+2\lambda+b_{P+2}\\ 2n+2\lambda+1, n+2\lambda+a_P\end{matrix}\middle|\frac{1}{w}\right),$$

where the upper (lower) signs are taken for $0\le\arg w<2\pi$ ($-2\pi<\arg w\le 0$). Furthermore, with $\sigma+\frac{1}{2}\neq$ an integer, and the w plane cut from 1 to $-\infty$ such that $\arg w=\arg(w-1)=0$ for $1<w$, we have for $|w-1|<1$,

(F)
$$e^{\pm i\pi(n+2\lambda)}g_n(we^{\pm i\pi})=e^{\pm i\pi(n+2\lambda)}g_n\left(e^{\pm i\pi}+(w-1)e^{\pm i\pi}\right)$$
$$=\frac{\Gamma(n+1)\Gamma(-\sigma-\frac{1}{2})}{\Gamma(n+2\lambda)}(w-1)^{\sigma+1/2}S_n^*(w)+R_n(w),$$

where $S_n^*(w)$, $R_n(w)$ are analytic functions of w in $|w-1|<1$, with $S_n^*(1)=1$.

Proof. For $|w|>1$, we can assume $g_n(w)$ and $l_{n,j}(w)$ are defined by the series expansions (0.1) and (0.2), respectively. As these series can be analytically continued through an arbitrary range of $\arg w$, provided $|w|>1$, (A) and (B) follow from elementary series manipulations. For $0<|w|<1$, consider the reflection formula identity

$$\Gamma(a_j-s)\Gamma(1-a_j+s)=e^{i2\pi(a_j-s)}\Gamma(a_j-s)\Gamma(1-a_j+s)+(-i2\pi)e^{i\pi(a_j-s)},$$

and the partial fraction identity, $y=e^{-i2\pi s}$,

$$(2\pi)^2e^{i\pi(1/2+2\gamma)}e^{-i2\pi s}\frac{\Gamma(a_P-s)\Gamma(1-a_P+s)}{\Gamma(b_{P+2}-s)\Gamma(1-b_{P+2}+s)}$$
$$=D_1\frac{\prod_{j=1}^{p+2}(y-e^{-i2\pi b_j})}{\prod_{j=1}^{p}(y-e^{-i2\pi a_j})}=\sum_{j=1}^{p}\frac{(i2\pi)C_jy}{ye^{i\pi a_j}-e^{-i\pi a_j}}-1+D_2y+D_1y^2$$
$$=\sum_{j=1}^{p}C_je^{-i\pi s}\Gamma(a_j-s)\Gamma(1-a_j+s)-1+D_2e^{-i2\pi s}+D_1e^{-i4\pi s}.$$

Multiply these two identities by

$$\frac{\Gamma(n+1)\Gamma(n+2\lambda+s)\Gamma(b_{P+2}-s)}{\Gamma(n+2\lambda)\Gamma(n+1-s)\Gamma(a_P-s)}(we^{i2\pi})^s$$

and integrate them along a contour L_+ which separates the poles of $\Gamma(b_{P+2}-s)$ from those of $\Gamma(n+2\lambda+s)\Gamma(1-a_P+s)$. The first integrated identity leads to (D), and the second to an expansion of

$$G_{p+2,\,p+2}^{0,\,p+1}\left(w\middle|\begin{matrix}1-n-2\lambda, a_P, n+1\\ b_{P+2}\end{matrix}\right)\equiv 0,$$

which reduces to (C). For further identities of this form, see Meijer's paper [8].

From the series definition in (0.2), it appears that $w=1$ is a singularity of $l_{n,j}(w)$. However, from the Mellin–Barnes integral for $l_{n,j}(w)$ in (0.2), it follows that $l_{n,j}(w)$ is analytic in the sector $|\arg w|<2\pi$, which includes the region $|w-1|<1$. Thus, $w=1$ is only an apparent singularity of $l_{n,j}(w)$.

Equation (E) follows from the analytic continuation of the series in (0.1), while (F) follows from some general results on hypergeometric functions. Nørlund in his classical analysis [9] of the hypergeometric equation, $\mathcal{H}y=0$,

$$\mathcal{H}=\prod_{j=1}^{q}(\delta+\beta_j-1)-z\prod_{j=1}^{q}(\delta+a_j), \qquad \delta=z\frac{d}{dz}, \quad \alpha_j, \beta_j \text{ complex parameters},$$

which is satisfied around $z=0$ by

$$z^{1-\beta_j}{}_{q+1}F_q\left(\begin{matrix}1,1-\beta_j+\alpha_Q\\ 1-\beta_j+\beta_Q\end{matrix}\middle| z\right), \qquad j=1,\cdots,q,$$

and around $z=\infty$ by

$$z^{-\alpha_j}{}_{q+1}F_q\left(\begin{matrix}1,1+\alpha_j-\beta_Q\\ 1+\alpha_j-\alpha_Q\end{matrix}\middle| \frac{1}{z}\right), \qquad j=1,\cdots,q,$$

showed that it is possible to choose a basis $\mathcal{B}=\{B_j(z): j=1,\cdots,q\}$ of $\mathcal{H}y=0$ in $|z-1|<1$ in such a way that it contains at most one element singular at $z=1$—say $B_1(z)$, while the remaining $q-1$ elements of $\mathcal{B}$ are analytic in $|z-1|<1$. In particular, for

$$\sigma_q=-1+\sum_{j=1}^{q}\beta_j-\sum_{j=1}^{q}\alpha_j \neq \text{ an integer},$$

Nørlund showed that we can take $B_1(z)$ to have the form

$$B_1(z)=(z-1)^{\sigma_q}S^*(z), \qquad S^*(1)=1,$$

where $S^*(z)$ is analytic in $|z-1|<1$. Expressing the solutions around $z=\infty$ in terms of the basis $\mathcal{B}$, he showed that

(2.22)

$$z^{-\alpha_j}{}_{q+1}F_q\left(\begin{matrix}1,\alpha_j+1-\beta_Q\\ \alpha_j+1-\alpha_Q\end{matrix}\middle| \frac{1}{z}\right)=\frac{\Gamma(\alpha_j+1-\alpha_Q)\Gamma(-\sigma_q)}{\Gamma(\alpha_j+1-\beta_Q)}(z-1)^{\sigma_q}S^*(z)+R_j(z),$$

$$\alpha_j-\alpha_Q,\ \alpha_j-\beta_Q \neq \text{ a negative integer},$$

where $R_j(z)$ is a solution of $\mathcal{H}_y=0$, analytic in $|z-1|<1$. With the identification

$$z^{-\alpha_q}{}_{q+1}F_q\left(\begin{matrix}1,\alpha_q+1-\beta_Q\\ \alpha_q+1-\alpha_Q\end{matrix}\middle| \frac{1}{z}\right)=w^{-n-2\lambda}{}_{p+2}F_{p+1}\left(\begin{matrix}n+2\lambda+b_{P+2}\\ 2n+2\lambda+1, n+2\lambda+a_P\end{matrix}\middle| \frac{1}{w}\right),$$

$$z=w, \quad j=q=p+2, \quad \alpha_q=n+2\lambda, \quad \alpha_{q-1}=-n,$$

$$1-\alpha_{Q-2}=a_P, \quad 1-\beta_Q=b_{P+2}, \quad \sigma_q=\sigma+\tfrac{1}{2},$$

(2.22) reduces to (F). Also, this shows $g_n(-w)$ and $l_{n,j}(w)$ satisfy $\mathcal{H}y=0$.

Remark 8. For $p=0$, Theorem 4(F) corresponds to

$$e^{\pm i\pi(n+2\lambda)}g_n(we^{\pm i\pi})$$
$$=\frac{\Gamma(n+1)\Gamma(n+2\lambda+b_1)\Gamma(n+2\lambda+b_2)\Gamma(1-2\lambda-b_1-b_2)}{\Gamma(n+2\lambda)\Gamma(n+1-b_1)\Gamma(n+1-b_2)}\,w^{b_2}$$
$$\times {}_2F_1\left(\begin{matrix}b_2-n, b_2+n+2\lambda\\ b_2+2\lambda+b_1\end{matrix}\middle|1-w\right)$$
$$+\frac{\Gamma(n+1)\Gamma(-1+2\lambda+b_1+b_2)}{\Gamma(n+2\lambda)}w^{b_1}(w-1)^{1-2\lambda-b_1-b_2}$$
$$\times {}_2F_1\left(\begin{matrix}-b_2-n+1-2\lambda, -b_2+n+1\\ -b_2-b_1+2-2\lambda\end{matrix}\middle|1-w\right), \qquad |w-1|<1.$$

3. A difference equation for $g_n(w)$ and $l_{n,j}(w)$.

THEOREM 5. *For $p+1$ a positive integer, $j,q=1,\cdots,p$ and $k=1,\cdots,p+2$, let*

(1) *$a_j-a_q\neq$ an integer $(j\neq q)$;*

(2) *$b_k-a_j\neq$ a negative integer.*

Then for m an integer, the functions $g_n(we^{i\pi 2m})$, $l_{n,j}(we^{i\pi(2m+1)})$, $j=1,\cdots,p$, satisfy the linear difference equation $\mathfrak{M}_n y_n=0$,

$$\mathfrak{M}_n=\prod_{j=1}^{p+2}\mathfrak{V}_n(2\lambda+1-j,b_j)-wn(n+2\lambda-2-p)\mathfrak{E}^{-1}\prod_{j=1}^{p}\mathfrak{V}_n(2\lambda+1-j,-1+a_j),$$

$$\prod_{j=1}^{r}P_j=P_rP_{r-1}\cdots P_1, \qquad \mathfrak{E}^{-j}y_n=y_{n-j},$$

$$\mathfrak{V}_n(\lambda,\mu)=\frac{(n+\lambda-1)(n-\mu)}{2n+\lambda-1}\mathfrak{E}^0-\frac{n(n+\lambda-1+\mu)}{2n+\lambda-1}\mathfrak{E}^{-1}.$$

Moreover, if r,s are integers, and w belongs to the nonempty region

$$\mathfrak{D}_{r,s}=\left\{w:\begin{matrix}\pi\,\mathrm{Max}\{-3-2r,-3-2s\}<\arg w<\pi\,\mathrm{Min}\{1-2r,1-2s\}\\ -2\pi<\arg(1+we^{i\pi 2s})<2\pi\end{matrix}\right\},$$

then the functions $l_{n,j}(we^{i\pi(2r+1)})$, $j=1,\cdots,p$, $g_n(we^{i\pi 2s})$, $g_n(we^{i\pi(2s+2)})$ form a basis $\mathfrak{B}_{r,s}$ of the difference equation $\mathfrak{M}_n y_n=0$, provided $g_n(we^{i\pi(2s+2)})$ is $g_n(we^{i\pi 2s})$ analytically continued along a curve $\mathcal{C}$ which encloses $w=0$, but not $w=-1$.

Proof. From the fact that

$$\mathfrak{V}_n(\lambda,\mu)\left\{\frac{(n+\lambda)_s}{(n+1)_{-s}}\right\}=\frac{(n+\lambda-1)_s}{(n+1)_{-s}}(s-\mu),$$

it follows that

(3.1)

$$\mathfrak{M}_n\left\{\frac{(n+2\lambda)_s}{(n+1)_{-s}}\right\}=\frac{(n+2\lambda-2-p)_s}{(n+1)_{-s}}\prod_{j=1}^{p+2}(s-b_j)$$
$$-w\frac{(n+2\lambda-2-p)_{s+1}}{(n+1)_{-s-1}}\prod_{j=1}^{p}(s+1-a_j).$$

Applying $\mathfrak{M}_n$ to the Mellin–Barnes integral representation for $g_n(we^{i\pi 2m})$, one obtains

$$I=\mathfrak{M}_n\{g_n(w^{i\pi 2m})\}$$
$$=\frac{(-1)^{p+2}}{2\pi i}\left(\int_L-\int_{1+L}\right)\frac{\Gamma(1+b_{P+2}-s)(n+2\lambda-2-p)_s}{\Gamma(a_P-s)(n+1)_{-s}}(we^{i\pi 2m})^s ds,$$

where the contour L is chosen to separate the poles of $\Gamma(b_{P+2}-s)$ from those of $\Gamma(n+2\lambda+s)$. I is equal to the sum of the residues between L and $1+L$. But by inspection, the integrand of I has no poles between L and $1+L$ if n is sufficiently large, so that $I=0$. The computation for $l_{n,j}(we^{i\pi(2m+1)})$ is similar.

Thus, the elements of $\mathfrak{B}_{r,s}$ are solutions of $\mathfrak{M}_n y_n=0$, and Theorems 2, 3, 4 imply that they are linearly independent as functions of n. Here, we use the fact that if $We^{i2\pi}$ is connected to W by $\mathcal{C}$, $W=we^{i\pi 2s}$,

$$\xi(We^{i2\pi})=\log\left\{e^{i\pi}\sqrt{W}+\sqrt{1+W}\right\}=-\xi(W).$$

Remark 9. $\mathfrak{M}_n$ can also be written in the form

$$\mathfrak{M}_n=\sum_{j=0}^{p+2}\{C_j(n,\lambda)+wD_j(n,\lambda)\}\mathcal{E}^{-j},\qquad D_0(n,\lambda)=D_{p+2}(n,\lambda)=0,$$

where the $C_j(n,\lambda)$, $D_j(n,\lambda)$ are rational functions of n and λ. It follows from (3.1) that with $q=p+2$,

$$\prod_{j=1}^{q}(s-b_j)=\sum_{j=0}^{q}C_j(n,\lambda)\frac{(s-n)_j(s+n+2\lambda-q)_{q-j}}{(-n)_j(n+2\lambda-q)_{q-j}},$$

$$(s-n)(s+n+2\lambda-q)\prod_{j=1}^{p}(s+1-a_j)=\sum_{j=0}^{q}D_j(n,\lambda)\frac{(s-n)_j(s+n+2\lambda-q)_{q-j}}{(-n)_j(n+2\lambda-q)_{q-j}}.$$

Explicit expressions for the $C_j(n,\lambda)$, $D_j(n,\lambda)$ can then be deduced from [4, Lemma 2.1] or [7, Vol. II, p. 139].

COROLLARY 5.1. *With the same conditions and notation as in Theorem* 5, *let*

$$b_m-b_k\neq \text{an integer}\ (m\neq k),\qquad m,k=1,\cdots,p+2.$$

For $|w|<1$, *the functions*

$$f_{n,k}(w)=\frac{(n+2\lambda)_{b_k}}{(n+1)_{-b_k}}w^{b_k}{}_{p+3}F_{p+2}\left(\begin{matrix}1,b_k-n,b_k+n+2\lambda,b_k+1-a_P\\ b_k+1-b_{P+2}\end{matrix}\middle|-w\right),$$
$$k=1,\cdots,p+2$$

form a basis $\mathfrak{B}_0$ *of* $\mathfrak{M}_n y_n=0$, *for n sufficiently large. For* $|w|>1$, *the functions*

$$m_{n,j}(w)=\frac{(n+2\lambda)_{-1+a_j}}{(n+1)_{1-a_j}}$$
$$\times w^{-1+a_j}{}_{p+3}F_{p+2}\left(\begin{matrix}1,-a_j+1+b_{P+2}\\ -a_j+n+2,-a_j+2-n-2\lambda,-a_j+1+a_P\end{matrix}\middle|\frac{-1}{w}\right),$$
$$n+2\lambda-1+a_j\neq \text{a positive integer},\quad j=1,\cdots,p,$$

$$g_n(w)=\frac{\Gamma(n+1)\Gamma(n+2\lambda+b_{P+2})w^{-n-2\lambda}}{\Gamma(n+2\lambda)\Gamma(2n+2\lambda+1)\Gamma(n+2\lambda+a_P)}$$

$$\times {}_{p+2}F_{p+1}\left(\begin{matrix} n+2\lambda+b_{P+2} \\ 2n+2\lambda+1, n+2\lambda+a_P \end{matrix}\middle| \frac{-1}{w}\right),$$

$$h_n(w)=\frac{\Gamma(n+1)\Gamma(2n+2\lambda)\Gamma(n+1-a_P)}{\Gamma(n+2\lambda)\Gamma(n+1-b_{P+2})}w^n\,{}_{p+2}F_{p+1}\left(\begin{matrix} -n+b_{P+2} \\ 1-2n-2\lambda, -n+a_P \end{matrix}\middle| \frac{-1}{w}\right),$$

$$2n+2\lambda, n+1-a_j \neq \textit{a positive integer}, \qquad j=1,\cdots,p,$$

form a basis $\mathfrak{B}_\infty$ *of* $\mathfrak{M}_n y_n=0$, *for n sufficiently large.*

Proof. Let the parameter γ be defined as in Theorem 4. As

$$f_{n,k}(w)=e^{-i\pi b_k}\frac{\Gamma(b_k+1-b_{P+2})}{\Gamma(b_k+1-a_P)}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}$$

$$\times G^{1,p+2}_{p+3,p+3}\left(we^{i\pi}\middle| \begin{matrix} 1-n-2\lambda, a_P, b_k, n+1 \\ b_k, b_{P+2} \end{matrix}\right),$$

it follows as in Theorem 5, that $\mathfrak{M}_n\{f_{n,k}(w)\}=0$. Then, from the explicit formulae

$$g_n(w)=\sum_{k=1}^{p+2}\frac{\Gamma(b^{*k}_{P+2}-b_k)}{\Gamma(a_P-b_k)}f_{n,k}(w), \tag{3.2}$$

$$l_{n,j}(we^{i\pi})=\sum_{k=1}^{p+2}\frac{\Gamma(b^{*k}_{P+2}-b_k)}{\Gamma(a^{*k}_P-b_k)}\Gamma(b_k+1-a_j)e^{i\pi b_k}f_{n,k}(w),$$

it follows that the $(p+2)$ functions in $\mathfrak{B}_0$ span the $(p+2)$-dimensional space spanned by $\mathfrak{B}_{r,s}$ in Theorem 5, and are linearly independent as functions of n. Note that (3.2) can be inverted as follows. From the partial fraction decomposition $y=e^{-i2\pi s}$,

$$\frac{\prod_{j=1,j\neq k}^{p+2}(y-e^{-i2\pi b_j})}{\prod_{j=1}^{p}(y-e^{-i2\pi a_j})}=y+d_0+\sum_{j=1}^{p}\frac{d_j y}{y-e^{-i2\pi a_j}},$$

it follows that

$$e^{i\pi s}\frac{\Gamma(a_P-s)\Gamma(1-a_P+s)}{\Gamma(b^{*k}_{P+2}-s)\Gamma(1-b^{*k}_{P+2}+s)}$$

$$=\frac{e^{-i\pi(b_k-2\gamma)}}{2\pi}+\frac{e^{i\pi(b_k-2\gamma)}}{2\pi}e^{i2\pi s}$$

$$+\sum_{j=1}^{p}\frac{\Gamma(a^{*j}_P-a_j)\Gamma(1-a_P+a_j)e^{i\pi s}\Gamma(a_j-s)\Gamma(1-a_j+s)}{\Gamma(b^{*k}_{P+2}-a_j)\Gamma(1-b^{*k}_{P+2}+a_j)}.$$

Multiplying this identity by

$$\frac{\Gamma(n+1)\Gamma(n+2\lambda+s)\Gamma(b_{P+2}-s)}{\Gamma(n+2\lambda)\Gamma(n+1-s)\Gamma(a_P-s)}w^s, \qquad |w|<1,$$

and integrating it over a contour L_+ which separates the poles of $\Gamma(b_{P+2}-s)$ from those of $\Gamma(n+2\lambda+s)\Gamma(1-a_P+s)$, we deduce that

$$(3.3)\qquad e^{i\pi b_k}\frac{\Gamma(b_k+1-a_P)}{\Gamma(b_k+1-b_{P+2})}f_{n,k}(w)$$

$$=\sum_{j=1}^{p}\frac{\Gamma\left(a_P^{*j}-a_j\right)\Gamma(1-a_P+a_j)}{\Gamma\left(b_{P+2}^{*k}-a_j\right)\Gamma\left(1-b_{P+2}^{*k}+a_j\right)}l_{n,j}(we^{i\pi})$$

$$+\frac{e^{-i\pi(b_k-2\gamma)}}{2\pi}g_n(w)+\frac{e^{i\pi(b_k-2\gamma)}}{2\pi}g_n(we^{i2\pi}).$$

This formula also serves to analytically extend $f_{n,k}(w)$ beyond $|w|<1$.

For $|w|>1$, we note that

$$(3.4)\qquad l_{n,j}(we^{i\pi})=\Gamma(n+2\lambda+a_j)$$

$$\times\Gamma(1-n-2\lambda-a_j)e^{-i\pi(n+2\lambda)}g_n(w)$$

$$+\frac{\Gamma(1-a_j+b_{P+2})}{\Gamma(1-a_j+a_P)}e^{i\pi(-1+a_j)}m_{n,j}(w).$$

As $\Gamma(n+a)\Gamma(1-n-a)e^{-i\pi n}$ is a periodic function of n whose period is equal to one, and $l_{n,j}(we^{i\pi})$, $g_n(w)$ are solutions of $\mathfrak{M}_n y_n=0$, so is $m_{n,j}(w)$ a solution of $\mathfrak{M}_n y_n=0$. To show that $h_n(w)$ is also a solution of $\mathfrak{M}_n y_n=0$, we could represent $h_n(w)$ as a G-function, say

$$h_n(w)=\frac{\Gamma(n+1)\Gamma(2n+2\lambda)\Gamma(1-2n-2\lambda)\Gamma(n+1-a_P)\Gamma(-n+a_P)}{\Gamma(n+2\lambda)\Gamma(n+1-b_{P+2})\Gamma(-n+b_{P+2})}$$

$$\times G_{p+2,p+2}^{p+2,1}\left(w\left|\begin{matrix}n+1,1-n-2\lambda,a_P\\ b_{P+2}\end{matrix}\right.\right),$$

and proceed as in Theorem 5, or expand this G-function representation by the residue calculus into a linear combination of the $f_{n,k}(w)$. However, the following derivation is more informative. From (3.2) with $|w|<1$, we have

$$(3.5)\qquad g_n(we^{i2\pi})=\sum_{k=1}^{p+2}e^{i2\pi b_k}\frac{\Gamma\left(b_{P+2}^{*k}-b_k\right)}{\Gamma(a_P-b_k)}f_{n,k}(w).$$

The G-function representation

$$f_{n,k}(w)=\frac{\Gamma(1+b_k-b_{P+2})\Gamma(n+1)}{\Gamma(1+b_k-a_P)\Gamma(1+n-b_k)\Gamma(b_k-n)\Gamma(n+2\lambda)}$$

$$\times G_{p+3,p+3}^{1,p+3}\left(w\left|\begin{matrix}1-n-2\lambda,a_P,n+1,b_k\\ b_k,b_{P+2}\end{matrix}\right.\right),$$

serves to analytically extend $f_{n,k}(w)$ into $|w|>1$, i.e.,

$$(3.6)\quad \frac{\Gamma(1+b_k-a_P)\Gamma(b_k-n)\Gamma(1+n-b_k)}{\Gamma(1+b_k-b_{P+2})} f_{n,k}(w)$$

$$= \sum_{j=1}^{p} \frac{\Gamma\left(a_j-a_P^{*j}\right)}{\Gamma\left(a_j-b_{P+2}^{*k}\right)} \Gamma(1-a_j+b_k)\Gamma(n+2-a_j)\Gamma(-n-1+a_j) m_{n,j}(w)$$

$$+ \frac{\Gamma(2n+2\lambda+1)\Gamma(-2n-2\lambda)\Gamma(n+2\lambda+a_P)\Gamma(1-n-2\lambda-a_P)}{\Gamma\left(n+2\lambda-b_{P+2}^{*k}\right)\Gamma\left(1-n-2\lambda-b_{P+2}^{*k}\right)} g_n(w)$$

$$+ \frac{\Gamma(2n+2\lambda)\Gamma(1-2n-2\lambda)\Gamma(1+n-a_P)\Gamma(-n+a_P)}{\Gamma\left(1+n-b_{P+2}^{*k}\right)\Gamma\left(-n+b_{P+2}^{*k}\right)} h_n(w).$$

Substituting (3.6) into (3.5) we obtain an expansion of $g_n(we^{i2\pi})$ in terms of $\mathfrak{B}_\infty$. To identify the connecting constants in this expansion in closed form, we assume $|w|<1$, and note that from the Mellin–Barnes integral expansion for $g_n(w)$, with $L=L_+$,

$$g_n(we^{i2\pi}) - e^{-i2\pi\sigma} g_n(w)$$

$$= (i2\pi) e^{-i\pi\sigma} \frac{\Gamma(n+1)}{\Gamma(n+2\lambda)} G_{p+3,p+3}^{p+2,1}\left(we^{i\pi} \,\middle|\, \begin{matrix} 1-n-2\lambda, a_P, n+1, 1-\sigma \\ b_{P+2}, 1-\sigma \end{matrix} \right),$$

σ an arbitrary complex parameter. Choosing $\sigma=n+2\lambda$, this reduces to

$$(3.7)\quad g_n(we^{i2\pi}) - e^{-i2\pi(n+2\lambda)} g_n(w)$$

$$= (i2\pi) e^{-i\pi(n+2\lambda)} \frac{\Gamma(n+1)}{\Gamma(n+2\lambda)} G_{p+2,p+2}^{p+2,0}\left(we^{i\pi} \,\middle|\, \begin{matrix} 1-n-2\lambda, a_P, n+1 \\ b_{P+2} \end{matrix} \right), \quad |w|<1.$$

Assume, for the moment, the following analytic continuation result.

LEMMA 8. *Let q be a positive integer, a_j, b_k be complex parameters such that*

$$a_j-a_k, b_j-b_k \neq \text{ an integer } (j\neq k), \qquad j,k=1,\cdots,q,$$

and the z-plane be cut from $-\infty$ to 0, and from 1 to ∞ such that $\arg z = \arg(1-z) = 0$ *for* $0<z<1$, *and* $(1-z)=e^{-i\pi}(z-1)$. *Then the analytic continuation with respect to z of*

$$G(z) = G_{q,q}^{q,0}\left(z \,\middle|\, \begin{matrix} a_Q \\ b_Q \end{matrix} \right) = \sum_{k=1}^{q} \frac{\Gamma\left(b_Q^{*k}-b_k\right)}{\Gamma(a_Q-b_k)} z^{b_k}\, {}_{q+1}F_q\left(\begin{matrix} 1, 1-a_Q+b_k \\ 1-b_Q+b_k \end{matrix} \,\middle|\, z \right), \qquad 0<z<1,$$

to the ray $z>1$, is given by

$$e^{\pm i\pi(1+\tau)} G^{\#}(z), \qquad \tau = \sum_{r=1}^{q} (b_r-a_r),$$

$$G^{\#}(z) = G_{q,q}^{0,q}\left(z \,\middle|\, \begin{matrix} a_Q \\ b_Q \end{matrix} \right) = \sum_{j=1}^{q} \frac{\Gamma\left(a_j-a_Q^{*j}\right)}{\Gamma(a_j-a_Q)} z^{-1+a_j}\, {}_{q+1}F_q\left(\begin{matrix} 1, 1+b_Q-a_j \\ 1+a_Q-a_j \end{matrix} \,\middle|\, \frac{1}{z} \right), \qquad 1<z,$$

where the upper (lower) sign is taken for the analytic continuation through the upper (lower) half plane.

Analytically continuing (3.7) with respect to w into $|w|>1$ through the upper half plane, and applying Lemma 8 to the resulting equation, we have for $|w|>1$

(3.8)

$$
\begin{aligned}
&g_n(we^{i2\pi}) \\
&\quad = e^{-i2\pi(n+2\lambda)}g_n(w)+(2\pi)e^{i\pi(2\gamma-n)}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)} \\
&\qquad\qquad \times G^{0,p+2}_{p+2,p+2}\left(we^{i\pi} \,\middle|\, \begin{matrix} 1-n-2\lambda, a_P, n+1 \\ b_{P+2} \end{matrix} \right) \\
&\quad = \sum_{j=1}^{p} (2\pi)e^{i\pi(2\gamma-n+a_j)}\frac{\Gamma(a_j-a_P^{*j})}{\Gamma(a_j-b_{P+2})}\Gamma(1+n-a_j)\Gamma(-n+a_j)m_{n,j}(w) \\
&\qquad + \Big\{ e^{-i2\pi(n+2\lambda)}+(2\pi)e^{i2\pi(\gamma-\lambda-n)} \\
&\qquad\qquad \times \frac{\Gamma(2n+2\lambda+1)\Gamma(-2n-2\lambda)\Gamma(n+2\lambda+a_P)\Gamma(1-n-2\lambda-a_P)}{\Gamma(n+2\lambda+b_{P+2})\Gamma(1-n-2\lambda-b_{P+2})} \Big\} g_n(w) \\
&\qquad + (2\pi)e^{i\pi 2\gamma}h_n(w).
\end{aligned}
$$

As the connecting constants in (3.8) are periodic functions of n whose period is 1, and $g_n(w)$, $g_n(we^{i2\pi})$ and the $m_{n,j}(w)$ are solutions of $\mathfrak{M}_n y_n=0$, it follows that $h_n(w)$ is a solution also. In addition, (3.4) and (3.8) imply that the $(p+2)$ functions in $\mathfrak{B}_\infty$ span the $(p+2)$-dimensional space spanned by $\mathfrak{B}_{r,s}$ in Theorem 5 and hence that the elements of $\mathfrak{B}_\infty$ are linearly independent as functions of n.

We now prove Lemma 8. Note that for $q=1$, the lemma reduces to the elementary result

$$
\frac{z^{b_1}(1-z)^{-1+a_1-b_1}}{\Gamma(a_1-b_1)} = e^{\pm i\pi(1+b_1-a_1)}\frac{z^{-1+a_1}(1-1/z)^{-1+a_1-b_1}}{\Gamma(a_1-b_1)},
$$

where the upper (lower) sign is taken for the analytic continuation through the upper (lower) half plane. For $q>1$, we proceed as follows. The series definition of $G(z)$ serves to analytically continue $G(z)$ into $|z|<1$, $|\arg z|<\pi$. Similarly, $e^{\pm i\pi(1+\tau)}G^{\#}(z)$ can be analytically continued into $|z|>1$, $|\arg(z-1)|<\pi$. Next, consider the Mellin–Barnes integral: $k=1,\cdots,q$,

$$
\begin{aligned}
H_k(ze^{-i\pi}) &= G^{1,q}_{q,q}\left(ze^{-i\pi} \,\middle|\, \begin{matrix} a_Q \\ b_k, b_Q^{*k} \end{matrix} \right) \\
&= \frac{1}{2\pi i}\int_{L_0} \frac{\Gamma(b_k-t)\Gamma(1-a_Q+t)}{\Gamma(1-b_Q^{*k}+t)}(ze^{-i\pi})^t dt,
\end{aligned}
$$

which is well defined for $|\arg(ze^{-i\pi})|<\pi$ or $0<\arg z<2\pi$. If the contour L_0 is deformed to the right into a contour L_+, we obtain from the residue calculus,

$$
H_k(ze^{-i\pi}) = \frac{\Gamma(1-a_Q+b_k)}{\Gamma(1-b_Q+b_k)}(ze^{-i\pi})^{b_k}{}_{q+1}F_q\left(\begin{matrix} 1, 1-a_Q+b_k \\ 1-b_Q+b_k \end{matrix} \,\middle|\, z \right), \tag{3.9}
$$

$$
|z|<1, \quad |\arg(ze^{-i\pi})|<\pi.
$$

Similarly, if L_0 is deformed to the left into a contour L_-, we obtain the result

(3.10)
$$H_k(ze^{-i\pi})=\sum_{j=1}^{q}\frac{\Gamma(a_j-a_Q^{*j})\Gamma(1+b_k-a_j)}{\Gamma(a_j-b_Q^{*k})}(ze^{-i\pi})^{-1+a_j}{}_{q+1}F_q\left(\begin{matrix}1,1+b_Q-a_j\\1+a_Q-a_j\end{matrix}\middle|\frac{1}{z}\right),$$
$$|z|>1,\quad |\arg(ze^{-i\pi})|<\pi.$$

Thus, the function $H_k(ze^{-i\pi})$ serves to analytically continue the series on the right-hand side of (3.9) through the upper half plane into the region $|z|>1$. Moreover, as $G(z)$ can be written as a linear combination of the $H_k(ze^{-i\pi})$, (3.10) serves to analytically continue $G(z)$ through the upper half plane into $|z|>1$, i.e.,

(3.11)
$$G(z)=\sum_{k=1}^{q}e^{i\pi b_k}\frac{\Gamma(b_Q^{*k}-b_k)\Gamma(1-b_Q^{*k}+b_k)}{\Gamma(a_Q-b_k)\Gamma(1-a_Q+b_k)}H_k(ze^{-i\pi})$$
$$=\sum_{j=1}^{q}C_j(ze^{-i\pi})^{-1+a_j}{}_{q+1}F_q\left(\begin{matrix}1,1+b_Q-a_j\\1+a_Q-a_j\end{matrix}\middle|\frac{1}{z}\right),$$
$$|z|>1,\quad |\arg(ze^{-i\pi})|<\pi,$$

where the constants C_j are given by

$$C_j=\frac{\Gamma(a_j-a_Q^{*j})}{\Gamma(a_j-b_Q)}\sum_{k=1}^{q}e^{i\pi b_k}\frac{\Gamma(b_Q^{*k}-b_k)\Gamma(1-b_Q^{*k}+b_k)}{\Gamma(1-a_Q^{*j}+b_k)\Gamma(a_Q^{*j}-b_k)}.$$

For an alternate representation of the C_j, consider, $y=e^{-i2\pi s}$,

$$T(y)=\frac{\prod_{k=1,k\neq j}^{q}(y-e^{-i2\pi a_k})}{\prod_{k=1}^{q}(y-e^{-i2\pi b_k})}=\sum_{k=1}^{q}\frac{T_k}{y-e^{-i2\pi b_k}}$$
$$=\sum_{k=1}^{q}\frac{T_k}{(2\pi i)}e^{i\pi(s+b_k)}\Gamma(b_k-s)\Gamma(1-b_k+s)$$
$$=\frac{e^{i\pi(s+a_j+\tau)}}{2\pi i}\cdot\frac{\Gamma(b_Q-s)\Gamma(1-b_Q+s)}{\Gamma(a_Q^{*j}-s)\Gamma(1-a_Q^{*j}+s)}.$$

Note that

$$-T(0)=\sum_{k=1}^{q}e^{i2\pi b_k}T_k=e^{i2\pi(\tau+a_j)},\qquad T_k=e^{i\pi(\tau-b_k+a_j)}\frac{\Gamma(b_Q^{*k}-b_k)\Gamma(1-b_Q^{*k}+b_k)}{\Gamma(a_Q^{*j}-b_k)\Gamma(1-a_Q^{*j}+b_k)},$$

so that

$$C_j=\frac{\Gamma(a_j-a_Q^{*j})}{\Gamma(a_j-b_Q)}e^{-i\pi(\tau+a_j)}\sum_{k=1}^{q}e^{i2\pi b_k}T_k=\frac{\Gamma(a_j-a_Q^{*j})}{\Gamma(a_j-b_Q)}e^{i\pi(\tau+a_j)}.$$

Substituting this representation of C_j into (3.11) and identifying the resulting series with the series continuation of $e^{i\pi(1+\tau)}G^{\#}(z)$, we readily arrive at Lemma 8. Similarly, to analytically continue $G(z)$ through the lower half plane, we consider $H_k(ze^{i\pi})$, and proceed as above.

For completeness, we note that if (3.4) is substituted into (3.8), and the resulting series summed as in the above lemma, we have

(3.12)

$$h_n(w)=\sum_{j=1}^{p} e^{-i\pi n}\frac{\Gamma(a_j-a_P^{*j})\Gamma(1-a_j+a_P)}{\Gamma(a_j-b_{P+2})\Gamma(1-a_j+b_{P+2})}\Gamma(1+n-a_j)l_{n,j}(we^{i\pi})$$
$$+\left\{\frac{e^{i2\pi(\gamma-n)}}{2\pi}\right.$$
$$\left.+e^{-i2\pi(n+\lambda)}\frac{\Gamma(2n+2\lambda)\Gamma(1-2n-2\lambda)\Gamma(n+1-a_P)\Gamma(-n+a_P)}{\Gamma(n+1-b_{P+2})\Gamma(-n+b_{P+2})}\right\}g_n(w)$$
$$+\frac{e^{-i2\pi\gamma}}{2\pi}g_n(we^{i2\pi}).$$

Remark 10. Using (3.3), (3.4) and (3.12), asymptotic expansions for the $f_{n,k}(w)$, $m_{n,j}(w)$ and $h_n(w)$ can be deduced from those for $g_n(w)$ and $l_{n,j}(w)$ in Theorems 2 and 3.

Remark 11. Should $2n+2\lambda$ take on a positive integer value, m say, all the elements of $\mathfrak{B}_\infty$ remain well defined, except for $h_n(w)$. A $(p+2)$nd linearly independent solution of $\mathfrak{M}_n y_n=0$ in $|w|>1$ is then given by, $2n+2\lambda=m-\varepsilon$,

$$h_n^*(w)=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left\{\frac{\Gamma(1+n-a_P)\Gamma(-n+a_P)}{\Gamma(1+n-b_{P+2})\Gamma(-n+b_{P+2})}g_n(w)-\varepsilon h_n(w)\right\}.$$

Similarly, one can deal with the other parameter restrictions.

4. The generalized Jacobi functions. With the identification

$$\begin{aligned}&\alpha_j=1-a_j, \quad j=1,\cdots,p,\\ &\beta_j=1-b_j, \quad j=1,\cdots,p+2, \qquad \beta_{p+2}=1 \text{ or } b_{p+2}=0,\end{aligned}\tag{4.1}$$

denote $g_n(w)$, $l_{n,j}(w)$, $m_{n,j}(w)$, etc. by $g_n^{\#}(w)$, $l_{n,j}^{\#}(w)$, $m_{n,j}^{\#}(w)$, etc., respectively. The generalized Jacobi functions $\mathcal{J}_n(w)$ are defined by

$$\mathcal{J}_n(w)=\frac{\Gamma(\beta_{P+1})\Gamma(n+1)}{\Gamma(\alpha_P)\Gamma(n+2\lambda)}G^{1,p+1}_{p+2,p+2}\left(w\,\middle|\,\begin{matrix}1-n-2\lambda\,1-\alpha_P,n+1\\ 0,1-\beta_{P+1}\end{matrix}\right)\tag{4.2}$$
$$={}_{p+2}F_{p+1}\left(\begin{matrix}-n,n+2\lambda,\alpha_P\\ \beta_{P+1}\end{matrix}\middle|w\right), \qquad |w|<1,$$

where n need not be an integer. If $n+1$ is a positive integer, $\mathcal{J}_n(w)$ is a polynomial.

Restating Theorem 2 and 3, we have the following.

THEOREM 6. *For $p+1$ a positive integer, $j,q=1,\cdots,p$ and $k=1,2,\cdots,p+2$, let $n,\lambda,\alpha_j,\beta_j$ be complex parameters such that n is large, $|n|\to+\infty$, $\arg n=\mathcal{O}(n^{-1})$ as $n\to\infty$, λ,α_j,β_j are bounded with respect to n, and*

$$\alpha_j-\alpha_q\neq \textit{an integer } (j\neq q),$$
$$\alpha_j-\beta_k\neq \textit{a negative integer}.$$

Also, let the region $\mathfrak{D}$ be defined by

$$\mathfrak{D}=\left\{w:\ \begin{array}{l}|\arg(w)|\le 2\pi-\varepsilon_1,\ |w|\ge|N|^{-1}\{\log|N|\}^{1+\varepsilon_3}\\ |\arg(1-w)|\le 2\pi-\varepsilon_2,\ |1-w|\ge|N|^{-1}\{\log|N|\}^{1+\varepsilon_4}\end{array}\right\},$$

where $N=\sqrt{n(n+2\lambda)}$, or $(n+\lambda)=\sqrt{N^2+\lambda^2}$, and the ε_j are small positive numbers independent of n.

Then for $m+1$ an arbitrary positive integer,

(A)
$$\frac{\Gamma(\alpha_j+1-\alpha_P)}{\Gamma(\alpha_j+1-\beta_{P+2})}\,l^{\#}_{n,j}(w)$$
$$=\sum_{k=0}^{m-1}(-1)^k\frac{(\alpha_j+1-\beta_{P+2})_k w^{-k-\alpha_j}}{(\alpha_j+n+1)_k(\alpha_j-n+1-2\lambda)_k(\alpha_j+1-\alpha_P)_k}+\mathcal{O}\left([wN^2]^{-m-\alpha_j}\right)$$
$$\sim\frac{(n+2\lambda)_{-\alpha_j}}{(n+1)_{\alpha_j}}w^{-\alpha_j}{}_{p+3}F_{p+2}\left(\begin{matrix}1,\alpha_j+1-\beta_{P+2}\\ \alpha_j+n+1,\alpha_j-n+1-2\lambda,\alpha_j+1-\alpha_P\end{matrix}\middle|\frac{1}{w}\right),$$
$$(N^2w)^{-1}=o(1),\quad N\to\infty,\quad |\arg w|\le 2\pi-\varepsilon_1,\quad j=1,\cdots,p,$$

and there exist functions $T^{\#}_k(\theta)$ such that

(B)
$$g^{\#}_n(we^{\pm i\pi})=\sqrt{\pi}\,[N^2w]^{\gamma}(1-w)^{-\gamma-\lambda}\exp\{\mp i2(n+\lambda)\theta\pm i\pi\gamma+T^{\#}(N,\pm\theta)\},$$
$$T^{\#}(N,\theta)=\sum_{k=1}^{m-1}T^{\#}_k(\theta)(iN)^{-k}+\left(\frac{\omega}{N}\right)^m\mathcal{O}(1),\qquad N\to\infty,\quad w\in\mathfrak{D},$$
$$2\omega=\frac{1+|w|}{\sqrt{|w(1-w)|}},\qquad 2\gamma=\frac{1}{2}+\sum_{j=1}^{p}\alpha_j-\sum_{j=1}^{p+1}\beta_j,\quad \sigma=-2\lambda-2\gamma,$$
$$\sqrt{w}=\sin\theta,\qquad \sqrt{1-w}=\cos\theta,\qquad 2\sqrt{w(1-w)}=\sin 2\theta,\qquad 1-2w=\cos 2\theta,$$
$$T^{\#}_k(\theta)=\sum_{j=0}^{[k/2]}U^{\#}_{k-2j,k}(\cot\theta)^{k-2j}+\sum_{j=0}^{[(k-1)/2]}V^{\#}_{k-2j,k}(\cot 2\theta)^{k-2j}=(-1)^kT^{\#}_k(-\theta),$$
$$U^{\#}_{11}=U_{11},\qquad U^{\#}_{22}=U_{22},\qquad 2U^{\#}_{02}=-2U_{02}+\lambda^2(\sigma+2\lambda),$$
$$U^{\#}_{33}=U_{33},\qquad 2U^{\#}_{13}=-2U_{13}+\lambda^2U_{11},$$
$$2V^{\#}_{11}=-4V^{\#}_{22}=\sigma(\sigma+1),\qquad 24V^{\#}_{33}=-\sigma(\sigma+1)(\sigma+3)(\sigma-2),$$
$$8V^{\#}_{13}=-\sigma(\sigma+1)\{2d_1-[(\sigma+\lambda)(\sigma+\lambda+1)+(\lambda^2+\lambda+2)]\},\ \textit{etc.},$$

where the $U_{k-2j,k}$ are defined in Corollary 2.1, *and the d_j are defined in* (1.3) *and Remark* 1, *using the identification in* (4.1). *Also, if $A(N,\theta)$, $P(N,\theta)$ are defined by*

$$T^{\#}(N,\pm\theta)=A(N,\theta)\pm iP(N,\theta),$$

then for $w\in\mathfrak{D}$,

$$A(N,\theta)=\frac{T^{\#}(N,\theta)+T^{\#}(N,-\theta)}{2}=\sum_{k=1}^{m-1}(-1)^kT^{\#}_{2k}(\theta)N^{-2k}+\left(\frac{\omega}{N}\right)^{2m}\mathcal{O}(1),\quad N\to\infty,$$

and

$$P(N,\theta)=\frac{T^{\#}(N,\theta)-T^{\#}(N,-\theta)}{2i}$$
$$=\sum_{k=0}^{m-1}(-1)^k T^{\#}_{2k+1}(\theta)N^{-2k-1}+\left(\frac{\omega}{N}\right)^{2m+1}\mathcal{O}(1),\qquad N\to\infty.$$

Proof. Everything follows in a straightforward manner from Theorems 2, 3 and Corollaries 2.1, 2.2. Note that the right-hand side of (A) is just $e^{\pm i\pi\alpha_j}m^{\#}_{n,j}(we^{\pm i\pi})$ treated as an asymptotic series for large n. For (B), we initially take $w\in(0,1)$, so that $\theta\in(0,\pi/2)$. Then

$$\xi(we^{\pm i\pi})=\log\left\{e^{\pm i\pi/2}\sqrt{w}+\sqrt{1-w}\right\}=\pm i\theta.$$

By analytic continuation, the relation $\xi(we^{i\pi})+\xi(we^{-i\pi})=0$ continues to hold for all $w\in\mathcal{D}$. The rest is by computation.

THEOREM 7. *With the notation of Theorem* 6, *let the parameters* α_j, β_k, λ *be independent of the large parameter* n, $|n|\to+\infty$, $\arg n=\mathcal{O}(n^{-1})$ *as* $n\to\infty$, *and satisfy the conditions,*

$$\alpha_j-\alpha_q\neq \text{an integer}\ (j\neq q),\qquad j,q=1,\cdots,p,$$
$$\beta_k-1\neq \text{a negative integer},\qquad k=1,\cdots,p+1.$$

Then for n sufficiently large,

(A)
$$\mathcal{J}_n(w)=\sum_{j=1}^{p}E_j l^{\#}_{n,j}(w)+F_1 g^{\#}_n(we^{-i\pi})+F_2 g^{\#}_n(we^{i\pi}),$$

$$E_j=\frac{\Gamma(\beta_{P+1})\Gamma(\alpha_j+1-\alpha_P)\Gamma\left(-\alpha_j+\alpha^{*j}_P\right)}{\Gamma(\alpha_P)\Gamma(\alpha_j+1-\beta_{P+1})\Gamma(-\alpha_j+\beta_{P+1})},$$

$$F_1=\frac{e^{i2\pi\gamma}\Gamma(\beta_{P+1})}{2\pi\Gamma(\alpha_P)},\qquad F_2=\frac{e^{-i2\pi\gamma}\Gamma(\beta_{P+1})}{2\pi\Gamma(\alpha_P)}.$$

If $n+1$ *is not a positive integer,* (A) *serves to analytically continue* $\mathcal{J}_n(w)$ *into* $|w|>1$.
For $|\arg w|\leq 2\pi-\varepsilon_1$, $(N^2w)^{-1}=o(1)$ *as* $N\to\infty$,

(B)
$$\sum_{j=1}^{p}E_j l^{\#}_{n,j}(w)\sim\sum_{j=1}^{p}\frac{\Gamma\left(-\alpha_j+\alpha^{*j}_P\right)\Gamma(\beta_{P+1})(n+2\lambda)_{-\alpha_j}}{\Gamma\left(\alpha^{*j}_P\right)\Gamma(-\alpha_j+\beta_{P+1})(n+1)_{\alpha_j}}w^{-\alpha_j}$$
$$\times{}_{p+3}F_{p+2}\left(\begin{matrix}1,\alpha_j+1-\beta_{P+2}\\ \alpha_j+n+1,\alpha_j-n+1-2\lambda,\alpha_j+1-\alpha_P\end{matrix}\middle|\frac{1}{w}\right).$$

For $|\arg w|\leq 2\pi-\varepsilon_1$, $|\arg(1-w)|\leq 2\pi-\varepsilon_2$,

(C)
$$F(\theta)=F_1 g^{\#}_n(we^{-i\pi})+F_2 g^{\#}_n(we^{i\pi})$$
$$=\frac{\Gamma(\beta_{P+1})}{\sqrt{\pi}\,\Gamma(\alpha_P)}[N^2w]^{\gamma}(1-w)^{-\gamma-\lambda}\exp\{A(N,\theta)\}\cos\{2(n+\lambda)\theta+\pi\gamma-P(N,\theta)\},$$

$$w\in\mathcal{D},\quad N\to\infty.$$

Combining (B) *and* (C), *we have the asymptotic expansion of* $\mathcal{J}_n(w)$ *for* $w\in\mathcal{D}$, $N\to\infty$. *More restricted versions of* (C) *are the following.*

For $|\arg(w-1)|\le\pi-\varepsilon_2$,

$$\text{(D)}\qquad F\Big(\frac{\pi}{2}+i\eta\Big)\sim\frac{\Gamma(\beta_{P+1})}{\sqrt{\pi}\,\Gamma(\alpha_P)}[N^2w]^{\gamma}(w-1)^{-\gamma-\lambda}$$

$$\times e^{-i\pi n+A(N,\pi/2+i\eta)}\cosh\Big[2(n+\lambda)\eta+iP\Big(N,\frac{\pi}{2}+i\eta\Big)\Big],$$

$$w\in\mathcal{D},\quad N\to\infty,\quad \cosh 2\eta=2w-1,\quad \eta>0\quad \textit{for}\ \arg(w-1)=0.$$

For $|\arg(we^{-i\pi})|\le\pi-\varepsilon_1$,

(E)

$$F(i\phi)\sim\frac{\Gamma(\beta_{P+1})}{\sqrt{\pi}\,\Gamma(\alpha_P)}[N^2we^{-i\pi}]^{\gamma}(1+we^{-i\pi})^{-\gamma-\lambda}e^{A(N,i\phi)}\cosh[2(n+\lambda)\phi+iP(N,i\phi)],$$

$$w\in\mathcal{D},\quad N\to\infty,\quad \cosh 2\phi=1+2we^{-i\pi},\quad \phi>0\quad \textit{for}\ \arg w=\pi.$$

Proof. (A) follows from (3.3) with $k=p+2$, and the identification in (4.1). (B) and (C) follow from Theorem 6. For (D), we note that

$$\cosh\eta=\sqrt{w},\qquad \sinh\eta=\sqrt{w-1},$$

and that $\operatorname{Re}(\eta)>0$ for $|\arg(w-1)\le\pi-\varepsilon_2$. Similarly, for (E) we have

$$\sinh\phi=\sqrt{we^{-i\pi}},\qquad \cosh\phi=\sqrt{1+we^{-i\pi}},$$

and $\operatorname{Re}\phi>0$ for $|\arg(we^{-i\pi})|\le\pi-\varepsilon_1$. Making use of the fact

$$e^a\cosh b=\cosh(a+b)+e^{-b}\sinh a,$$

(D) and (E) follow directly from (C).

Remark 12. The generalized Jacobi polynomials, $\mathcal{J}_n(w)$ with $n+1$ a positive integer, were treated from a differential equation viewpoint in [1], and from a generating function approach using Darboux's method in [2]. The present treatment is more general.

5. Asymptotic expansions for $g_n(e^{\pm i\pi})$, and $\mathcal{J}_n(1)$. Using the general theory of hypergeometric functions, it was shown in Theorem 4(F), that the principal branch of $g_n(w)$ has the following behavior near $w=e^{\pm i\pi}$:

$$(5.1)\quad e^{\pm i\pi(n+2\lambda)}g_n(we^{\pm i\pi})=e^{\pm i\pi(n+2\lambda)}g_n\big(e^{\pm i\pi}+(w-1)e^{\pm i\pi}\big),$$

$$=\frac{\Gamma(n+1)\Gamma(-\sigma-1/2)}{\Gamma(n+2\lambda)}(w-1)^{\sigma+1/2}S_n^*(w)+R_n(w),$$

$$\sigma+\frac{1}{2}=1-2\lambda+\sum_{k=1}^{p}a_k-\sum_{k=1}^{p+2}b_k\neq\text{ an integer},$$

$$|w-1|<1,\quad |\arg w|<\pi,\quad |\arg(w-1)|<\pi,$$

where $S_n^*(w)$, $R_n(w)$ are analytic in $|w-1|<1$, with $S_n^*(1)=1$. It can be shown that when $\sigma+\frac{1}{2}$ is an integer, similar expansions exist which involve $\log(w-1)$. From these

expansions, it follows that

$$g_n(e^{\pm i\pi}) = e^{\mp i\pi(n+2\lambda)} R_n(1) \text{ exists}, \quad \text{when } \mathrm{Re}\left(\sigma+\frac{1}{2}\right)>0. \tag{5.2}$$

For $w=\sinh^2\xi$, $\mathrm{Re}\,\xi\geq 0$, the asymptotic expansion of $g_n(w)$, as derived in Theorem 2, was based on the integral representation (1.13), which we now write in the unnormalized form,

$$g_n(w) = \sqrt{\pi}\,\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\,\frac{\Gamma(-\sigma)}{2\pi i}\int_{\mathcal{C}_0} e^{-2(\xi-u)(n+\lambda)} F(u,\xi)\,d(2u), \tag{5.3}$$

$$\sigma = \frac{1}{2} - 2\lambda + \sum_{k=1}^{p} a_k - \sum_{k=1}^{p+2} b_k \neq \text{an integer},$$

$$\begin{aligned} F(u,\xi) &= (\sinh\xi)^{-2\lambda-\sigma}(\cosh\xi)^{\sigma}(2u)^{\sigma}K(u,\xi) \\ &= [\sinh(\xi-u)]^{-2\lambda-2\sigma}\{(\sinh u)[\sinh(2\xi-u)]\}^{\sigma}\mathbb{S}^*(X), \\ &\qquad u\in\mathcal{C}_0, \quad \xi\in\mathcal{D}_\xi\cap\{\xi:\mathrm{Re}\,\xi\geq 0\}, \end{aligned} \tag{5.4}$$

where $K(u,\xi)$, $X, \mathbb{S}^*(X)$ are as in Corollary 1.1, and where $\mathcal{C}_0$ is the contour shown in Figs. 2 and 8. As each of the component, multiple-valued functions in the definition of $F(u,\xi)$ is well defined by its principal value when $u\in\mathcal{C}_0$, $\xi\in\mathcal{D}_\xi^0$—see Corollary 1.1, $F(u,\xi)$ and these component functions remain well defined when they are given a common analytic continuation with respect to ξ from $\mathcal{D}_\xi^0$ to $\mathcal{D}_\xi$. In particular, the behaviour of $F(u,\xi)$ for u near 0 follows directly from Lemma 2, where it is shown that $K(u,\xi)=1+\mathcal{O}(u)$, $u\to 0$, $\xi\in\mathcal{D}_\xi$. Note that the parameter σ is the same in both (5.1) and (5.3). Note also, that when $w\to e^{\pm i\pi}$, $\xi=\alpha+i\beta\to\pm i\pi/2$, and that the singularities of $[\sinh(2\xi-u)]^{\sigma}$ coalesce with those of $(\sinh u)^{\sigma}$, so that the previous analysis of (5.3) must be modified.

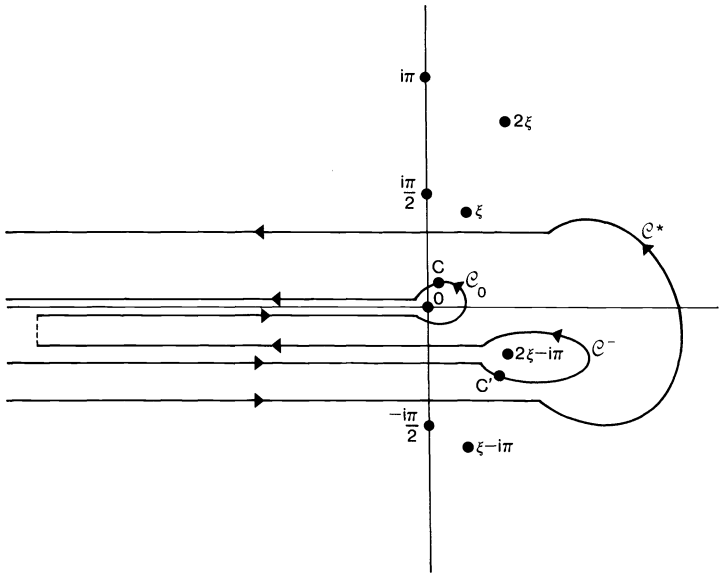


FIG 8. *u-surface.*

THEOREM 8. *Let $p,n,\lambda,a_j,b_j,\sigma,d_j$ be as in Theorem* 1, *with the additional restrictions that n be a large parameter such that $|n|\to+\infty$. $\arg n=\mathcal{O}(1)$ as $n\to\infty$, and λ,a_j,b_j be bounded with respect to n. Then for $\mathrm{Re}(\sigma+\frac{1}{2})>0$, there exist constants H_{2k} such that for $m+1$ an arbitrary positive integer,*

$$g_n(e^{\pm i\pi})$$

$$=\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}G^{p+2,1}_{p+2,p+2}\left(e^{\pm i\pi}\,\middle|\,\begin{matrix}1-n-2\lambda,a_P,n+1\\ b_{P+2}\end{matrix}\right)$$

$$=\frac{\Gamma(n+1)\Gamma(n+2\lambda+b_{P+2})e^{\mp i\pi(n+2\lambda)}}{\Gamma(n+2\lambda)\Gamma(2n+2\lambda+1)\Gamma(n+2\lambda+a_P)}{}_{p+2}F_{p+1}\left(\begin{matrix}n+2\lambda+b_{P+2}\\ 2n+2\lambda+1,n+2\lambda+a_P\end{matrix}\,\middle|\,1\right)$$

$$=e^{\mp i\pi(n+2\lambda)}\Gamma\left(\sigma+\frac{1}{2}\right)(n+\lambda)^{-2\sigma-2\lambda}H(n+\lambda),$$

$$H(n+\lambda)=1+\sum_{k=1}^{m-1}H_{2k}(n+\lambda)^{-2k}+(n+\lambda)^{-2m}\mathcal{O}(1),\qquad n+\lambda\to\infty,$$

$$2\gamma=-\frac{1}{2}+\sum_{j=1}^{p+2}b_j-\sum_{j=1}^{p}a_j,\qquad \sigma=-2\lambda-2\gamma,$$

$$6H_2=3(2\sigma+1)d_1-(\sigma+1)(2\sigma+1)(2\sigma+3\lambda)+\lambda(\lambda-1)(2\lambda-1),$$

$$\begin{aligned}1440H_4=&\,360(2\sigma+1)(2\sigma+3)d_2\\&+120(2\sigma+1)[\lambda(\lambda-1)(2\lambda-1)-(\sigma+2)(2\sigma+3)(2\sigma+3\lambda+2)]d_1\\&+(\sigma+1)(\sigma+2)(2\sigma+1)(2\sigma+3)[80\sigma^2+8\sigma(30\lambda+7)+15\lambda(6\lambda+1)]\\&-40(\sigma+1)(2\sigma+1)(2\sigma+3\lambda)\lambda(\lambda-1)(2\lambda-1)\\&+4\lambda(\lambda^2-1)(4\lambda^2-1)(5\lambda-6),\ \textit{etc.}\end{aligned}$$

Proof. We first prove Theorem 8 under the more restrictive condition

$$2\sigma\neq\text{ an integer.}\tag{5.5}$$

In what follows, let $\xi=\alpha+i\beta$ satisfy the conditions $\alpha>0$, $0\le\beta\le\pi/2$, and let $q_n(w)$, $h_n(w)$ be the functions which result when $\mathcal{C}_0$ in (5.3) is replaced by the contours $\mathcal{C}^-$, $\mathcal{C}^*$ in Fig. 8, respectively. Here, $\mathcal{C}^-$ is the contour $\mathcal{C}_0$ translated through $2\xi-i\pi$, and the contour $\mathcal{C}^*$ is a positively oriented loop contour which encloses $u=0$ and $u=2\xi-i\pi$, but which encloses none of the other singularities of $F(u,\xi)$, i.e., the points $u=k\xi+i\pi q$, $k=0,1$ or 2, q an integer. The initial (terminal) straight line segment of $\mathcal{C}^*$ is taken to coincide with the initial (terminal) straight line segment of $\mathcal{C}^-$ ($\mathcal{C}_0$), respectively. As $F(u,\xi)$ is well defined on this terminal straight line segment of $\mathcal{C}_0$, the definition of $F(u,\xi)$ can be extended to $\mathcal{C}^*$, and hence $\mathcal{C}^-$, by an analytic continuation with respect to u along these contours. Finally, as $\mathcal{C}^*$ can be deformed into $\mathcal{C}_0+\mathcal{C}^-$, without crossing any singularities of $F(u,\xi)$, it follows that

$$\begin{aligned}&h_n(w)=g_n(w)+q_n(w),\\&w=[\sinh\xi]^2,\qquad \xi=\alpha+i\beta,\quad 0<\alpha,\quad 0\le\beta\le\frac{\pi}{2}.\end{aligned}\tag{5.6}$$

Consider the integral for $q_n(w)$. For $u\in\mathcal{C}^-$, we can make the substitution $u=2\xi-i\pi+v$, $v\in\mathcal{C}_0$. By carefully tracing out the variation in argument over $\mathcal{C}^*$, we see that for $u\in\mathcal{C}^-$ and $v\in\mathcal{C}_0$,

$$\sinh u=e^{-i\pi}\sinh(2\xi+v)=\sinh(2\xi e^{-i\pi}-v),$$
$$\sinh(\xi-u)=\sinh(\xi+v)=e^{i\pi}\sinh(\xi e^{-i\pi}-v),$$
$$\sinh(2\xi-u)=\sinh v,$$
$$X=\left\{\frac{\sinh(\xi-u)}{\sinh\xi}\right\}^{-2}=\left\{\frac{\sinh(\xi e^{-i\pi}-v)}{\sinh(\xi e^{-i\pi})}\right\}^{-2},$$
$$X-1=\frac{(\sinh u)[\sinh(2\xi-u)]}{[\sinh(\xi-u)]^2}=e^{-i2\pi}\frac{[\sinh(2\xi e^{-i\pi}-v)](\sinh v)}{[\sinh(\xi e^{-i\pi}-v)]^2},$$

$$e^{-2(\xi-u)(n+\lambda)}F(u,\xi)=e^{-i2\pi(n+2\lambda+\sigma)}e^{-2(\xi e^{-i\pi}-v)}F(v,\xi e^{-i\pi}). \tag{5.7}$$

To see that the same branch of $\mathcal{S}^*(X)$ has been used on both sides of (5.7), consider the variation in argument of $X=X(u,\xi)$, and $X-1=(X-1)(u,\xi)$ as u and ξ vary. With reference to Fig. 8, let ξ be as above, but small, i.e., ≈ 0, and let C be the point on $\mathcal{C}_0$ where $\arg u=\arg\xi$, while C' is taken to be the point on $\mathcal{C}^-$ where $\arg(u-2\xi+i\pi)=\arg(\xi e^{-i\pi})$. Then for u near C on $\mathcal{C}_0$, $|(X-1)(u,\xi)|<1$,

$$\arg X(u,\xi)\approx 0,\qquad \arg(X-1)(u,\xi)\approx\arg u-\arg\xi\approx 0,$$

so that $\mathcal{S}^*(X(u,\xi))$ reduces to $\mathcal{S}_p^*(X(u,\xi))$, while for $u=2\xi-i\pi+v$ near C' on $\mathcal{C}^-$, $|(X-1)(u,\xi)|=|(X-1)(u,\xi e^{-i\pi})|<1$,

$$\arg X(u,\xi)=\arg X(v,\xi e^{-i\pi})\approx 0,$$
$$\arg(X-1)(u,\xi)=-2\pi+\arg(X-1)(v,\xi e^{-i\pi})$$
$$\approx -2\pi+\arg v-\arg(\xi e^{-i\pi})\approx -2\pi.$$

From (1.3), we note that for $|X-1|<1$, $|\arg X|<\pi$,

$$\mathcal{S}_p^*(X)=\mathcal{S}_p^*(1+e^{i2\pi q}(X-1)),\qquad q \text{ an integer},$$

so that $\mathcal{S}_p^*(X)$ can be analytically continued outside of $|\arg(X-1)|<\pi$ to an arbitrary sector of $\arg(X-1)$. Thus, the branch of $\mathcal{S}^*(X)$ obtained by analytically continuing $\mathcal{S}_p^*(X)$ from a neighborhood of C on $\mathcal{C}_0$ to a neighborhood of C' on $\mathcal{C}^-$, reduces, by the monodromy theorem, to just

$$\mathcal{S}^*(X(u,\xi))=\mathcal{S}_p^*(1+e^{-i2\pi}(X-1)(v,\xi e^{-i\pi}))=\mathcal{S}_p^*(X(v,\xi e^{-i\pi})).$$

By analytic continuation with respect to u along $\mathcal{C}^-$, it follows that the branch of $\mathcal{S}^*(X(u,\xi))$ occurring on the left-hand side of (5.7) agrees with the branch of $\mathcal{S}^*(X)$ which occurs in the definition of $F(v,\xi e^{-i\pi})$ as given by (5.4).

Substituting (5.7) into the integral for $q_n(w)$, and noticing that (5.3) with contour $\mathcal{C}_0$, is valid not only for ξ with $\mathrm{Re}\,\xi\geq 0$, but also for $\xi e^{-i\pi}$, provided ξ is sufficiently small, we can relate the resulting integral to $g_n(we^{-i2\pi})$, i.e.,

$$e^{i2\pi(n+2\lambda+\sigma)}q_n(w)=g_n([\sinh(\xi e^{-i\pi})]^2)=g_n(we^{-i2\pi}). \tag{5.8}$$

By analytic continuation, (5.8) continues to hold even when ξ is not small. In particular, when $\alpha>0$ and β is near $\pi/2$ so that $|w|=\sinh^2\alpha+\sin^2\beta>1$, we can use Theorem 4(A) to write (5.8) in the form

$$q_n(w)=e^{-i2\pi(n+2\lambda+\sigma)}g_n(we^{-i2\pi})=e^{-i2\pi\sigma}g_n(w), \qquad |w|>1. \tag{5.9}$$

Combining (5.6) and (5.9), we have the functional relationships,

$$h_n(w)=[1+e^{-i2\pi\sigma}]g_n(w), \quad g_n(w)=\frac{e^{i\pi\sigma}}{2\cos\pi\sigma}h_n(w), \qquad |w|>1.$$

Replacing $h_n(w)$ by its integral representation, we can write, after some simplification,

$$g_n(w)=e^{i\pi\sigma}4^{\sigma}\frac{\Gamma(\sigma+\frac{1}{2})\Gamma(n+1)}{\Gamma(n+2\lambda)}\frac{\Gamma(-2\sigma)}{2\pi i}\int_{\mathcal{C}^*}e^{-2(\xi-u)(n+\lambda)}F(u,\xi)\,d(2u), \tag{5.10}$$

$$0<\alpha, \quad 0\le\beta\le\sqrt{\tfrac{\pi}{2}}.$$

Again carefully tracing out the variation in arguments, as $\xi\to i\pi/2$ and $\mathcal{C}^*\to\mathcal{C}_0$, we see that for $u\in\mathcal{C}_0$,

$$\sinh(\xi-u)\to e^{i\pi/2}\cosh u, \qquad \sinh(2\xi-u)\to\sinh u,$$
$$e^{-2(\xi-u)}F(u,\xi)\to e^{-i\pi(n+2\lambda+\sigma)}e^{2u(n+\lambda)}(\cosh u)^{-2\lambda-2\sigma}(\sinh u)^{2\sigma}\mathbb{S}_p^*([\cosh u]^{-2}).$$

Applying this limit process to (5.10), we deduce the basic results

$$g_n(e^{i\pi})=e^{-i\pi(n+2\lambda)}\Gamma\left(\sigma+\frac{1}{2}\right)(n+\lambda)^{-2\sigma-2\lambda}H(n+\lambda), \tag{5.11}$$

$$H(n+\lambda)=(n+\lambda)^{2\sigma+2\lambda}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}\frac{\Gamma(-2\sigma)}{2\pi i}\int_{\mathcal{C}_0}e^{2u(n+\lambda)}(2u)^{2\sigma}K(u)\,d(2u), \tag{5.12}$$

$$K(u)=(\cosh u)^{-2\lambda}\left(\frac{\tanh u}{u}\right)^{2\sigma}\mathbb{S}_p^*([\cosh u]^{-2}), \qquad u\in\mathcal{C}_0.$$

Each of the component multiple-valued functions in the definition of $K(u)$ is defined by its principal value when $u\in\mathcal{C}_0$. From (5.2), it is clear that (5.11) remains valid if $(i\pi)$ is replaced by $(-i\pi)$. Note that $H(n+\lambda)$, as defined by (5.12), is well defined for all valued of σ, including $\mathrm{Re}(\sigma+\frac{1}{2})\le 0$, although limiting values need to be taken when 2σ is an integer. The asymptotic expansion of $H(n+\lambda)$ follows directly from the following lemmas.

LEMMA 9. *Let* $X=(\cosh u)^{-2}$ *so that* $1-X=(\tanh u)^2$, *and let* $K(u)$ *be as in* (5.12). *Then*

$$K(u)=X^{\lambda}\left(\frac{1-X}{u^2}\right)^{\sigma}\mathbb{S}_p^*(X), \qquad u\in\mathcal{C}_0,$$

can be analytically continued as a function of u *into a neighborhood of* $u=0$, *and, in particular, has the Maclaurin expansion*

$$K(u)=\sum_{k=0}^{\infty}(2u)^{2k}U_{2k}, \qquad U_0=1, \quad |u|<\frac{\pi}{4}.$$

Moreover, if $m+1$ is a positive integer, and

$$K(u)=\sum_{k=0}^{m-1}(2u)^{2k}U_{2k}+(2u)^{2m}K_{2m}(u),$$

then there exist proper numbers M,s (i.e., nonnegative and independent of n and u), such that

$$|K_{2m}(u)|\leq Me^{s|u|}, \qquad u\in\mathcal{C}_0.$$

Proof. Clearly, $(\cosh u)^{-2\lambda}\ [(\tanh u)/u]^{2\sigma}$ is analytic in $|u|<\pi/2$, while $\mathfrak{S}_p^*([\cosh u]^{-2})$ is an analytic function of u in $|\tanh u|\leq\tan|u|<1$, or $|u|<\pi/4$. The remainder of the lemma is proved in the same way Lemma 6 was proved.

LEMMA 10. *Let n be as above, and $H(n+\lambda)$ be defined by* (5.12). *Then for all values of σ, there exist constants H_{2k} such that for $m+1$ a positive integer,*

$$H(n+\lambda)=1+\sum_{k=1}^{m-1}H_{2k}(n+\lambda)^{-2k}+(n+\lambda)^{-2m}\mathcal{O}(1), \qquad n+\lambda\to\infty.$$

The H_{2k} here are the same as those in the statement of Theorem 8.

Proof. First, assume that 2σ is not an integer. Substituting the finite expansion for $K(u)$ from Lemma 9 into (5.12), and using the integral preceding (2.8), we have

$$H(n+\lambda)=(n+\lambda)^{2\lambda-1}\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)} \times\left\{\sum_{k=0}^{m-1}(2\sigma+1)_{2k}U_{2k}(n+\lambda)^{-2k}+H_{2m}(n+\lambda)\right\}, \tag{5.13}$$

$$H_{2m}(n+\lambda)=(n+\lambda)^{2\sigma+1}\frac{\Gamma(-2\sigma)}{2\pi i}\int_{-\infty}^{0+}e^{2u(n+\lambda)}(2u)^{2\sigma+2m}K_{2m}(u)\,d(2u). \tag{5.14}$$

Using the estimate for $K_{2m}(u)$ in Lemma 9, together with the change of variable $v=2u(n+\lambda)$, it is easy to see that

$$H_{2m}(n+\lambda)=(n+\lambda)^{-2m}\mathcal{O}(1), \qquad n+\lambda\to\infty. \tag{5.15}$$

The asymptotic expansion of $H(n+\lambda)$ then follows from (5.13), (5.15) and Lemma 7, with

$$H_{2k}=\sum_{j=0}^{k}e_{2j}(\lambda)(2\sigma+1)_{2k-2j}U_{2k-2j},$$

where the $e_{2j}(\lambda)$ are defined in Lemma 7. This establishes the lemma when 2σ is not an integer. An alternate representation for $H(n+\lambda)$ can be derived as follows. Choose m_0 such that $\mathrm{Re}(2\sigma+2m)>0$, when $m\geq m_0$. As $K_{2m}(u)$ takes on the same value at corresponding points on the integration contour in (5.14), e.g., $u=xe^{-i\pi}$ and $u=xe^{i\pi}$, all factors in the integrand of (5.14) are single-valued, except for $(2u)^{2\sigma+2m}$. Taking account of the branches of this factor, we can write

$$H_{2m}(n+\lambda)=\frac{(n+\lambda)^{2\sigma+1}}{\Gamma(2\sigma+1)}\int_0^{\infty}e^{-2v(n+\lambda)}(2v)^{2\sigma+2m}K_{2m}(ve^{\pm i\pi})\,d(2v), \qquad m\geq m_0. \tag{5.16}$$

This representation of $H_{2m}(n+\lambda)$ remains well defined when 2σ takes on integer values, although a limiting form of $K_{2m}(ve^{\pm i\pi})/\Gamma(2\sigma+1)$ has to be taken when 2σ is a negative integer. As the first m terms on the right of (5.13) remain well defined when 2σ is an integer, $H(n+\lambda)$ is well defined for all σ. Finally, as Lemma 9 remains valid when $\mathfrak{S}_p^*(X)$ is replaced by $\mathfrak{S}_p^*(X)/\Gamma(2\sigma+1)$ and 2σ takes on integer values, the change of variable $x=2v(n+\lambda)$ in (5.16) leads one immediately to (5.15), again, when $m\geq m_0$. This restriction on m_0 can be dropped, by noticing that

$$(2\sigma+1)_{2k}U_{2k}(n+\lambda)^{-2k}=(n+\lambda)^{-2k}\mathcal{O}(1), \qquad n+\lambda\to\infty,$$

for $k=0,1,\cdots,m_0-1$. This completes the proof of the lemma and the theorem.

Remark 13. If $g_n(we^{i\pi})$ and $g_n(we^{-i\pi})$ are the functions represented in (5.1), then

(5.17)

$$\begin{aligned} D_n(w)&=e^{i\pi(\sigma+2\lambda)}g_n(we^{i\pi})+e^{-i\pi(\sigma+2\lambda)}g_n(we^{-i\pi}) \\ &=(i2\pi)e^{i\pi\sigma}\frac{\Gamma(-\sigma-1/2)}{\Gamma(-n)\Gamma(n+2\lambda)}(w-1)^{\sigma+1/2}S_n^*(w)+2[\cos\pi(\sigma-n)]R_n(w). \end{aligned}$$

Thus, if $n+1$ is a positive integer, the singular term in (5.17) vanishes, and $D_n(1)$ is well defined for all σ. In particular,

$$\begin{aligned} (5.18)\qquad D_n(1)&=2[\cos\pi(\sigma-n)]R_n(1) \quad \text{if } \mathrm{Re}\left(\sigma+\frac{1}{2}\right)>0 \\ &=2[\cos\pi\sigma](-1)^nR_n(1) \text{ if } n+1 \text{ is a positive integer.} \end{aligned}$$

Remark 14. From (5.1) and (5.11), it follows that

$$(5.19)\qquad R_n(1)=\Gamma\left(\sigma+\frac{1}{2}\right)(n+\lambda)^{-2\sigma-2\lambda}H(n+\lambda),$$

when $\mathrm{Re}(\sigma+\frac{1}{2})>0$. As $R_n(1)$ and $H(n+\lambda)$ are well defined for all σ, and in fact, depend analytically on σ, (5.19) remains valid for all σ. The asymptotic expansion of $D_n(1)$ then follows from (5.19) and Lemma 10, for the cases listed in (5.18).

THEOREM 9. *Let*

$$\mathcal{J}_n(w)={}_{p+2}F_{p+1}\left(\begin{matrix}-n,n+2\lambda,\alpha_P \\ \beta_{P+1}\end{matrix}\middle| w\right), \qquad |w|<1, \quad \beta_{p+2}=1,$$

where $p+1$ is a positive integer, $n,\lambda,\alpha_j,\beta_j$ are complex parameters such that n is large, $|n|\to\infty$, $\arg n=\mathcal{O}(n^{-1})$ as $n\to\infty$, λ,α_j,β_j are bounded with respect to n, and

$$\begin{aligned} &\alpha_j-\alpha_k\neq \textit{an integer } (j\neq k), \qquad j,k=1,\cdots,p, \\ &\beta_k-1\neq \textit{a negative integer}, \qquad k=1,\cdots,p+1. \end{aligned}$$

Then $\mathcal{J}_n(1)$ exists as a well defined, finite, number, if

$$\begin{aligned} (A)\qquad &\mathrm{Re}\left(\sigma+\frac{1}{2}\right)=\mathrm{Re}\left(-2\lambda-2\gamma+\frac{1}{2}\right)>0, \quad \textit{or } n+1 \textit{ is a positive integer}, \\ &2\gamma=\frac{1}{2}+\sum_{j=1}^{p}\alpha_j-\sum_{j=1}^{p+1}\beta_j, \qquad \sigma=-2\lambda-2\gamma, \end{aligned}$$

and in fact,

$$\text{(B)}\qquad \mathcal{J}_n(1)=\sum_{j=1}^{p} E_j l_{n,j}^{\#}(1)+F_1 g_n^{\#}(e^{-i\pi})+F_2 g_n^{\#}(e^{i\pi}),$$

in the notation of Theorem 7. *Moreover,*

$$\text{(C)}\qquad \sum_{j=1}^{p} E_j l_{n,j}^{\#}(1)$$

$$\sim \sum_{j=1}^{p} \frac{\Gamma(-\alpha_j+\alpha_P^{*j})\Gamma(\beta_{P+1})(n+\lambda)_{-\alpha_j}}{\Gamma(\alpha_P^{*j})\Gamma(-\alpha_j+\beta_{P+1})(n+1)_{\alpha_j}}$$

$$\times {}_{p+3}F_{p+2}\left(\begin{matrix}1,\alpha_j+1-\beta_{P+2}\\ \alpha_j+n+1,\alpha_j-n+1-2\lambda,\alpha_j+1-\alpha_P\end{matrix}\middle|\,1\right),$$

as $n+\lambda\to\infty$, *and if*

$$\text{(D)}\qquad D_n^{\#}(1)=F_1 g_n^{\#}(e^{-i\pi})+F_2 g_n^{\#}(e^{i\pi})$$

$$=\frac{\Gamma(\beta_{P+1})}{\Gamma(\alpha_P)}\,\frac{\Gamma(\sigma+\frac{1}{2})(n+\lambda)^{4\gamma+2\lambda}}{\Gamma(\sigma+\frac{1}{2}-n)\Gamma(-\sigma+\frac{1}{2}+n)}H^{\#}(n+\lambda),$$

then there exists constants $H_{2k}^{\#}$ *such that for* $m+1$ *an arbitrary positive integer,*

$$H^{\#}(n+\lambda)=1+\sum_{k=1}^{m-1} H_{2k}^{\#}(n+\lambda)^{-2k}+(n+\lambda)^{-2m}\mathcal{O}(1),\qquad n+\lambda\to\infty,$$

where the $H_{2k}^{\#}$ *are the* H_{2k} *in Theorem* 8, *with the identification in* (4.1). *In particular, when* $n+1$ *is a positive integer,*

$$\frac{\Gamma(\sigma+\frac{1}{2})}{\Gamma(\sigma+\frac{1}{2}-n)\Gamma(-\sigma+\frac{1}{2}+n)}=\frac{(-1)^n}{\Gamma(\frac{1}{2}-\sigma)}=\frac{(-1)^n}{\Gamma(\frac{1}{2}+2\gamma+2\lambda)}.$$

Proof. The $l_{n,j}^{\#}(w)$ are analytic at $w=1$, and $D_n^{\#}(1)$ is a constant multiple of $D_n(w)$ at $w=1$ with the identification (4.1). (B) and (C) follow directly from Theorem 7(A) and (B), respectively. (D) follows from Theorem 8, (5.18) and (5.19).

Remark 15. Just as in Corollary 2.2, the asymptotic results in Theorem 8 and Theorem 9 (D) can be rewritten in terms of the large variable $N, N^2=n(n+2\lambda)$, or can be rewritten in exponential form.

REFERENCES

[1] J. L. FIELDS AND Y. L. LUKE, *Asymptotic expansions of a class of hypergeometric polynomials, with respect to the order*, J. Math. Anal. Appl., 6 (1963), pp. 394–403.

[2] J. L. FIELDS, *Asymptotic expansions of a class of hypergeometric polynomials with respect to the order*-III, J. Math. Anal. Appl., 12 (1965), pp. 593–601.

[3] _____, *A note on the asymptotic expansion of a ratio of gamma functions*, Proc. Edinburg Math. Soc., (2), 15 (1966), pp. 43–45.

[4] J. L. FIELDS, Y. L. LUKE AND J. WIMP, *Recursion formulae for generalized hypergeometric functions*, J. Approx. Theory, 1 (1968), pp. 137–166.

[5] J. L. FIELDS, *A linear scheme for rational approximations*, J. Approx. Theory, 6 (1972), pp. 161–175.

[6] _____, *Uniform asymptotic expansions of certain classes of Meijer G-functions for a large parameter*, this Journal, 4 (1973), pp. 482–507.

[7] Y. L. LUKE, *The Special Functions and Their Approximations*, Vol. I and II, Academic Press, New York, 1969.

[8] C. S. MEIJER, *On the G-function*, Proc. Kon. Ned. Akad. Wet., Amsterdam 49 (1946), pp. 227–237, 344–356, 457–469, 632–641, 765–772, 936–943, 1063–1072, 1165–1175.

[9] N. E. NØRLUND, *Hypergeometric functions*, Acta Math., 94 (1955), pp. 289–349.

[10] G. SZEGŐ, *Orthogonal Polynomials*, AMS Colloquium Publications, 23, American Mathematical Society, Providence, RI, 1959.

[11] G. N. WATSON, *Asymptotic expansions of hypergeometric functions*, Trans. Cambr. Phil. Soc., 22 (1918), pp. 277–308.